LT-EDI 2025

Fifth Workshop on Language Technology for Equality, Diversity, Inclusion

Proceedings of the Workshop

September 9, 2025

The LT-EDI organizers gratefully acknowledge the support from the following sponsors.

In cooperation with

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN None

Introduction

We are excited to welcome you to the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion (LT-EDI-2025), the 5th Conference on Language, Data and Knowledge (LDK). This year, the workshop will be held in a hybrid format (both online and Workshops will take place at Palazzo del Mediterraneo on 9th September 2025, while the main venue for the conference will be Palazzo Corigliano, on 10th - 11th September 2025, located in the Naples, Italy. With the rapid advancement of technology, digital communication has become a central part of daily life. While many globally dominant languages have successfully transitioned into the digital era, numerous regional and low-resource languages continue to face significant technological challenges. Equality, Diversity and Inclusion (EDI) is an important agenda across every field throughout the world. Language as a major part of communication should be inclusive and treat everyone with equality. Today's large internet community uses language technology (LT) and has a direct impact on people across the globe. EDI is crucial to ensure everyone is valued and included, so it is necessary to build LT that serves this purpose. Recent results have shown that big data and deep learning are entrenching existing biases and that some algorithms are even naturally biased due to problems such as 'regression to the mode'. Our focus is on creating LT that will be more inclusive of gender, racial, sexual orientation, persons with disability. The workshop will focus on creating speech and language technology to address EDI not only in English, but also in less resourced languages. The workshop received a total of 40 active submissions. Reviewer recruitment was highly effective, with 232 out of 249 invited reviewers accepting the invitation. Of the 270 assigned reviews, 117 were completed, resulting in a review submission rate of 43.33%. Additionally, 41.67% of reviewers (100 out of 240) completed all their assigned reviews. A majority of submissions (65%, or 26 out of 40) received at least three reviews, ensuring a robust evaluation process. Decisions were finalized for all submissions (100%), leading to an acceptance rate of 95% (38 papers). This included 6 papers (15%) accepted for oral presentations and 32 papers (80%) accepted for poster presentations. Only 2 submissions (5%) were rejected. There were no withdrawn submissions, and only one paper was desk rejected. These metrics reflect a thorough and inclusive review process, driven by active reviewer participation and a strong commitment to quality.

Program Committee

Program Chairs

Bharathi B
Paul Buitelaar, University of Galway
Bharathi Raja Chakravarthi, University of Galway
Shunmuga Priya Muthusamy Chinnan
Thenmozhi Durairaj
Miguel Ángel García
Salud María Jiménez-Zafra, Universidad de Jaén
Prasanna Kumar Kumaresan, Data Science Institution, University of Galway, Ireland
Rahul Ponnusamy

Reviewers

Habiba A, Keerthi Vasan A, Kritika A, Lekhashree A, Mohan Raj M A, Moogambigai A, Belo Abhigyan, Babatunde Abimbola Abiola, Tolulope Olalekan Abiola, Jidan Al Abrar, Tewodros Achamaleh, Raksha Adyanthaya, Shamima Afroz, Sabik Aftahee, Nitisha Aggarwal, Saurabh Aggarwal, N.Nasurudeen Ahamed, Anisha Ahmed, Kawsar Ahmed, S Ananthasivan, Mahfuz Ahmed Anik, K Anishka, Shuang Ao, Mohammad Shamsul Arefin, Priyanka Ashokan

Bharathi B, Premjith B, Tofayel Ahmmed Babu, Girma Yohannis Bade, Priyatharshan Balachandran, Nitin Nikamanth Appiah Balaji, Arupa Barua, Kavin Bharathi, J Bhuvana, Manan Buddhadev, Miriam Butt

Bagavathi C, Hare Ram C, Jerin Mahibha C, Dola Chakraborty, Trina Chakraborty, Soham Chaudhuri, Md. Sajid Alam Chowdhury, Minhaz Chowdhury

Arun Prasad T D, Pandiarajan D, Sakkthi Gurru D, Dipankar Das, Sayan Das, Sayan Das, Udoy Das, Somsubhra De, Ashraf Deen, Naihao Deng, Ashim Dey, MC Dhanush, Srijita Dhar, Thenmozhi Durairaj

Boomika E, Enjamamul Haque Eram

Wahid Faisal

Dhanyashree G, Gnanasabesan G, Jyothish Lal G, Payal Godhani, A. Justin Gopinath, Anusha M D Gowda, Pranav Gupta, V Gurucharan

Nida Hafeez, Adeep Hande, Tareque Md Hanif, Fariha Haq, Bachu Naga Sri Harini, Md Mehedi Hasan, Asha Hegde, Md. Sajjad Hossain, Md. Refaj Hossan

Ariful Islam

Pavithra J, Vikash J, Amit Jaspal, Abirami Jayaraman, Jayanth Jeyadevaswamy, Deeptanshu Jha, Sandra Johnson, Jobin Jose, Uma Jothi, Fiona Victoria Stanley Jothiraj

Arivuchudar K, Durai Singh K, Kalpana K, Nithish Ariyha K, Sitara K, Sreeja K, Srihari V K,

Md Minhazul Kabir, Santhosh Kakarla, Shreyas Karthik, Arunaggiri Pandian Karunanidhi, Dondluru Keerthana, Mahir Absar Khan, Rohith Gowtham Kodali, Olga Kolesnikova, Sai Koneru, Avaneesh Koushik, Mikhail Krasitskii, Mahankali Sri Ram Krishna, Abhinav Kumar

Aravindh M, Mithun M, Niranjan Kumar M, Geetha M P, Bitan Mallik, Dr G Manikandan, Poojitha Sai Manikandan, Durga Prasad Manukonda, Kankipati Venkata Meghana, Md Ayon Mia, Md. Alam Miah, Sabrina Afroz Mitu, Md. Mohiuddin, Hasan Murad, Jahnavi Murali

Hari Krishnan N, Radha N, Sripriya N, Gladiss Merlin N.r, Abdullah Al Nahian, Md. Mubasshir Naib, Hamada Nayel, Nishanth.S Nishanth.S, Aathavan Nithiyananthan, Ahamed Rameez Mohamed Nizzad, Keerthana Nnl, Syeda Alisha Noor

Temitope Oladepo

Lahari P, Lalith Kishore V P, Sarumathi P, Yeshwanth Balaji A P, Ashraful Islam Paran, Sarbajeet Pattanaik, Fred Philippy, Ravi Teja Potla, Rahatun Nesa Priti

Arthi R, Eshwanth Karti T R, Ponsubash Raj R, Ramesh Kannan R, Shankari S R, Swathika R, Vijay Manickam R, Mostafa Rahgouy, Abdur Rahman, MD.Mahadi Rahman, Md Mizanur Rahman, Sheikh Ayatur Rahman, Burugu Rahul, Mohan Raj, Ratnavel Rajalakshmi, Riya Rajeev, RajeswariRajasekar RajeswariRajasekar, RajeswariRajasekar RajeswariRajasekar, Kasu Sai Kartheek Reddy, Shruthi Rengarajan, Billodal Roy

Angel Deborah S, Dr Jannath Nisha O S, Ganesh Sundhar S, Meenakshy S, Sachin Kumar S, Vishal A S, B Saathvik, Dipanjan Saha, Bommineni Sahitya, Meetesh Saini, Nazmus Sakib, Tara Samiksha, Eric SanJuan, Minoru Sasaki, Dharunika Sasikumar, Khadiza Sultana Sayma, Siddhaarth Sekar, Aishwarya Selvamurugan, Diya Seshan, Aruna Devi Shanmugam, Kogilavani Shanmugavadivel, Anik Mahmud Shanto, Deepawali Sharma, Harshita Sharma, Rasha Sharma, Hosahalli Lakshmaiah Shashirekha, Md. Tanvir Ahammed Shawon, Arya Palackal Shijish, Gersome Shimi, Symom Hossain Shohan, Simran Simran, Aakash Singh, Abhai Pratap Singh, Vivek Kumar Singh, Vrijendra Singh, Bhuvaneswari Sivagnanam, Janeshvar Sivakumar, Rajalakshmi Sivanaiah, Raj Sonani, Ippatapu Venkata Srichandra, Tanisha Sriram, Subhashini Sudhakar, Sivasuthan Sukumar, Monorama Swain, Swetha.N.G Swetha.N.G

Radhika K T, Luxshan Thavarasa, Shanmitha Thirumoorthy, Farjana Alam Tofa, Ashutosh Tripathi

Mugilkrishna D U

Shruthikaa V, Vijay Karthick Vaidyanathan, Vajratiya Vajrobol, Adarsh Valoor, Advait Vats, Venkatesh Velugubantla, Satya Subrahmanya Gautama Shastry Bulusu Venkata, Konkimalla Laxmi Vignesh, Abhay Vishwakarma

Azmine Toushik Wasi, Sidney Wong

Abhishek Singh Yadav, Ashok Yadav, Mesay Gemeda Yigezu

Keynote Talk

Understanding Attention in Asymmetric Kernel Point of View

Dr. Soman K. P.
Amrita Vishwa Vidyapeetham, India

2025-05-03 09:15 – Room: **Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico, USA**

Abstract: Transformers has redefined deep learning research and has become the most prominent architecture across domains such as natural language processing, computer vision, and image processing. Attention mechanism, particularly self-attention, is central to the success of this architecture, which allows the model to capture dependencies across the input sequences. However, the fundamental challenge in understanding self-attention is its intrinsic symmetry. The existing works often consider self-attention as a kernel method, leveraging symmetric kernels based on Mercer's theorem. However, the self-attention matrices used in the transformer architectures are inherently asymmetric, which leads to an inconsistency between the theoretical formulation and the practical implementation. The primal-attention, a novel attention mechanism based on kernel singular value decomposition explicitly models the asymmetry. Therefore, reformulating self-attention using primal-dual representation ensures efficient computation and low-rank approximation that enhances performance and generalization.

Bio: Dr. Soman K. P. is the Dean of the School of Artificial Intelligence and Head of the Department at Amrita Vishwa Vidyapeetham, Coimbatore. With over 27 years of experience in research and teaching, his expertise spans Artificial Intelligence and Data Science. He has published more than 500 papers in leading journals and conferences, including IEEE Transactions, IEEE Access, and Applied Energy. He is the author of four books, including Insight into Wavelets, Insight into Data Mining (also translated into Chinese), Support Vector Machines and Other Kernel Methods, and Signal and Image Processing—the Sparse Way. Dr. Soman is the most cited researcher with over 10,000 citations. He has consistently been ranked among the world's top 2% most influential scientists by Stanford University for the past three years. His contributions have also been recognized by the Government of India and organizations like Springer Nature and Career 360. At CEN, he leads M.Tech programs in Computational Engineering and Networking (Data Science) and Computer Science and Engineering (Artificial Intelligence). A new B.Tech program in AI and Data Science launched under his leadership in 2023. He has guided over 20 Ph.D. scholars and currently supervises 8+ ongoing doctoral researchers. His current research interests include AI for DNA sequence analysis, reinforcement learning in robotics, computer vision, and cyber-physical systems.

Table of Contents

SSNCSE@LT-EDI-2025:Detecting Misogyny Memes using Pretrained Deep Learning models

Sreeja K, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
sreeja2350625@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Misogyny meme detection is identifying memes that are harmful or offensive to women. These memes can hide hate behind jokes or images, making them difficult to identify. It's important to detect them for a safer and respectful internet for everyone. Our model proposed a multimodal method for misogyny meme detection in Chinese social media by combining both textual and visual aspects of memes. The training and evaluation data were part of a shared task on detecting misogynistic content. We used a pretrained ResNet-50 architecture to extract visual representations of the memes and processed the meme transcriptions with BERT. The model fused modality-specific representations with a feed-forward neural net for classification. The selected pretrained models were frozen to avoid overfitting and to enhance generalization across all classes, and only the final classifier was fine-tuned on labelled meme recollection. The model was trained and evaluated using test data to achieve a macro F1-score of 0.70345. As a result, we have validated lightweight combining approaches for multimodal fusion techniques on noisy social media and how they can be validated in the context of hostile meme detection tasks.

1 Introduction

The rise of misogynistic content on social media contexts is increasingly problematic, particularly as social media platforms adopt more multimodal formats such as memes that combine text and images to propagate problematic, exclusionary, or unfair messages. In the context of multilingual and multicultural online spaces that include languages such as Chinese, Tamil, Malayalam, and Hindi-English code-mixed communities, identifying misuse introduces additional challenges associated with language, culture, and multimodal resources.

Recently, researchers have reported the difficulties in identifying misogynistic memes. The authors, (Lu et al., 2024) introduced the ToxiCN-MM dataset, the first large-scale collection of harmful memes in Chinese, and proposed a Multimodal Knowledge Enhanced (MKE) model tailored for culture-specific meme detection. Their findings demonstrate the complicated nature of identification models and consider the value of contextual and cultural knowledge in detection models. Similarly, shared tasks and resources such as the (Chakravarthi et al., 2024) and (Pattanaik et al., 2025) have taken this research into low-resource Dravidian languages such as Tamil and Malayalam, collecting and collaborating on annotated datasets such as MDMD, which advance the development of systems that can identify code-mixed, as well as monolingual, data.

The MIMIC project also released a large set of human-annotated examples that will detect misogyny in Hindi-English code-mixed memes and can further multimodal hate speech research in minority languages. These two projects show not only the need for effective image-text fusion-based models but also the need to consider cultural aspects of misogyny and multimodal aspects of the data when attempting to detect misogyny in online discourse.

The rest of the paper is organized as follows: Section 2 analyzes the related works done in the previous research, and Section 3 discusses the misogyny meme corpus in the current work. Section 4 contains a detailed discussion of the proposed models used in the current work. Section 5 explains the experimental results. Section 6 discusses the limitations. In Section 7 concludes the paper.

2 Related work

The arrival of social media has brought awareness to the issue of misogyny in multimodal content, including memes.(Chakravarthi et al., 2025)provides the overview of the Misogyny Meme Detection Shared Task for Chinese Social Media. The

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 1–5

task of detecting misogyny in multimodal content with the MAMI dataset was handled (Singh et al., 2023), who found that multimodal models combining BERT for text and Vision Transformer (ViT) for images, when pretrained on hate speech data, substantially outperformed unimodal models. As noted here, (Liu et al., 2019) introduced RoBERTa and specifically focused on building on BERT by reviewing the positive aspect of pretraining on a significant amount of large-scale data and longer in both iterations and time. In that evaluation, RoBERTa beat BERT in a variety of language tasks, reinforcing the trend of using pretrained models, frozen as feature extractors. In a similar vein, (Deng et al., 2022) proposed the COLD benchmark to identify offensive language in Chinese and established the claim that Singh et al. (2023)BERT-based models could identify nuance, including biases related to race, gender, and region. As multimodal approaches matured, the use of pretrained models for both textual and visual understanding became increasingly common. Memotion 2.0 was introduced in a recent study (Ramamoorthy et al., 2022), and prior studies showed improved meme sentiment and emotion classification by combining ResNet-50 for visual features and BERT for textual features. A recent contribution to the field (Gasparini et al., 2022) developed a benchmark dataset for meme-based misogyny detection (both direct and indirect) and emphasized the importance of expert annotations in overcoming sociocultural barriers and interpreting hostile or ironic content. Another study (Chakravarthi et al., 2024) illustrated that lightweight classifiers such as MLPs are effective when used with frozen feature extractors in multilingual meme classification.

While much of the research has focused on English and Indian languages, there is limited work on misogyny detection in multimodal Chinese-language content. This study addresses that gap by establishing a pipeline that combines ResNet-50 and BERT, both used in frozen mode, with an MLP classifier—achieving a macro-averaged F1-score of 0.70345. This demonstrates the feasibility of developing a robust misogyny detection model for Chinese social media using pretrained models and lightweight classification.

This study expands on previous work by applying multimodal detection of misogyny to Chinese, a linguistically and culturally important domain that has rarely been addressed in existing literature. Many previous studies were primarily based on fine-tuning of pretrained models. In this study, we employed frozen pretrained models (BERT and ResNet-50) along with a minimal MLP classifier, to demonstrate that there's an effective and efficient usage of pretrained models in this, and many other, domains. Additionally, our strong performance in identifying misogyny in Chinese memes indicates that multimodal pretrained are generalizable to domains beyond English and Indian languages. This provides more breadth and scalability to online multilingual safety prevention research.

3 Dataset Description

The task aims to develop a model for Misogyny Meme detection. The dataset for the Shared Task on Misogyny Meme Detection found at LT-EDI@LDK 2025(Ponnusamy et al., 2024), (Chakravarthi et al., 2024) has been designed to support multimodal and multilingual research on Chinese social media data. The dataset is comprised of memes, which consist of both an image and an accompanying textual component that are typically captured from an overlay caption or a comment historically associated with the image. The image contains a variety of meme formats shown on online forums. Each of the memes is assisted by a binary label, "Misogynistic" or "Non-misogynistic," which conveys whether the meme generates or expresses any form of gender hatred or bias. The dataset consists of three (3) datasets for training, development, and testing. The label distribution for training and development data is mentioned in Fig1 and Fig2.

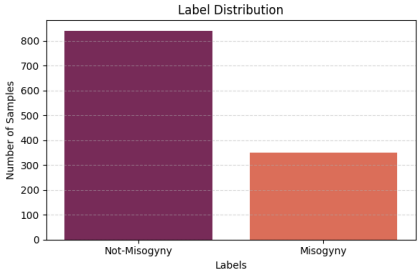

Figure 1: Training Data distribution

4 Proposed Methodology

This study presents a structured multimodal meme content classification approach into misogynistic and non-misogynistic classes. The study involves five basic steps: data preprocessing, feature extraction, model building, training, and prediction. Each step is well designed to address the unique chal-

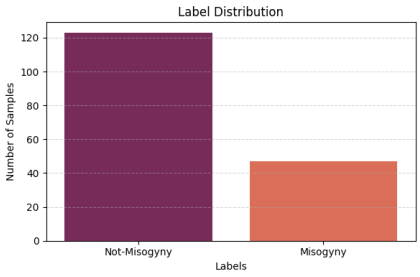

Figure 2: Development data distribution

lenges of integrating visual and textual information on low-context social media images.

4.1 Data Preprocessing

It ensures consistency and tidiness on both image and text modalities. Images are resized to a standard dimension of 224×224 and normalized via ImageNet mean and standard deviation values. Text data in the form of meme transcriptions is cleaned and tokenized with the pretrained BERT tokenizer. A PyTorch dataset class is employed to refine this process, ensuring alignment between text and image inputs and efficiently handling both training and test data formats.

4.2 Feature Extraction

Pretrained models are leveraged based on their stability to produce rich, high-level representations. The ResNet-50 model handles visual modality. It is pretrained on the ImageNet dataset. This model is frozen to prevent overfitting and preserve common visual features. At the same time, transcriptions are passed through a pretrained BERT-base model, and only the [CLS] token representation (pooler output) as the sentence-level embedding is used. Both ResNet-50 and BERT remain untrained to reduce computational costs and utilize the power of large-scale pretraining.

4.3 Model Architecture

It combines both modalities by concatenating the 2048-dimensional feature vector for images with 768-dimensional text embeddings to create a 2816-dimensional feature vector for multimodal representation. This vector is categorized using a lightweight feedforward classifier made of a fully connected layer, activation of ReLU, dropout regularization, followed by a classification layer that outputs logits for the binary classification as misogynist or not-misogynist.

4.3.1 Architecture Workflow

The architecture works in the following steps:

1. **Input Processing**
   Meme images and their corresponding transcriptions are taken as input.

2. **Visual Feature Extraction**
   The image is passed through a frozen ResNet-50 model, which gives a high-level visual feature representation of size 2048.

3. **Text Feature Extraction**
   The text is tokenized and passed through a frozen BERT-base model. The [CLS] token embedding is used as a 768-dimensional text representation.

4. **Feature Fusion**
   The image and text features are combined by simple concatenation to form a single 2816-dimensional feature vector.

5. **Classification**
   This fused feature is passed through a small feedforward neural network (MLP) that includes:

   - A fully connected layer with ReLU activation
   - Dropout for regularization
   - A final layer that predicts if the meme is misogynistic or not

4.4 Training Phase

The classifier parameters are trained while the pretrained backbone models are frozen. The model is trained using the Adam optimizer and a cross-entropy loss function. We use a batch size of 8 and a learning rate 1e-4 for five training epochs. The goal of this targeted training is to limit the risk of overfitting and also to allow the model to converge faster.

4.5 Prediction

The trained model will be used to predict unseen test data. Since the test samples now have contextual textual information associated with them, we will use both the images and their text in the prediction process. The model will interpret this multi-input to predict the probability of misogyny of any given sample. The predictions will be saved in a CSV file, with the predicted labels.

In summary, this methodology can provide an efficient and scalable pipeline for multimodal classification through the integration of a pretrained visual and language model with a selection of effective training and modularity, Multimodal features can be successfully leveraged while computationally efficient.

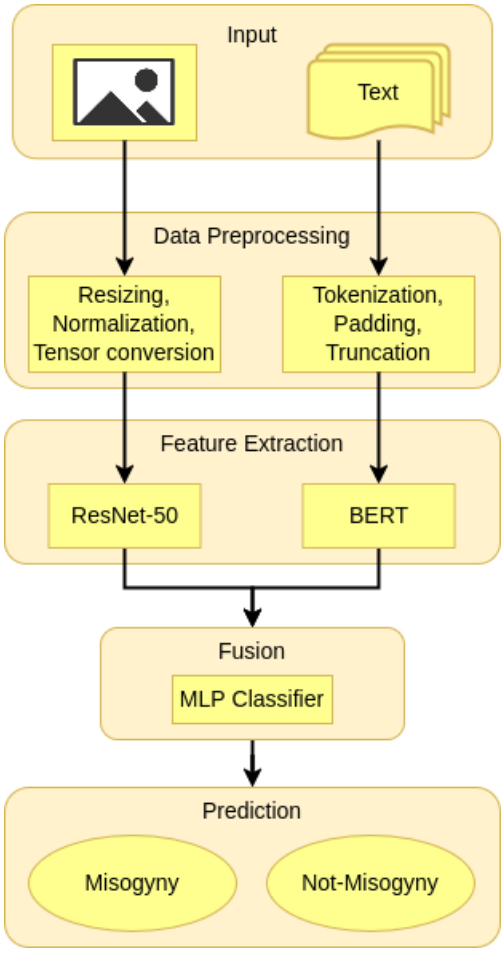

Figure 3: Architecture Diagram of Proposed Work

5 Experimental result

The proposed multimodal misogyny classification model, using both visual and text features, using ResNet50 and BERT respectively, shows that the model is learning sufficiently over the training periods of interest. The loss steadily decreases over each of the epochs, which indicates sufficient learning or convergence; for example, the training loss drops from approximately 0.55 in the first epoch to approximately 0.37 in the last epoch.

To assess the generalizability of the model, performance was evaluated on a development dataset, as shown in Table 1. The evaluation used familiar classification metrics of accuracy, precision, recall, and F1-score. The classifier achieved a macro-averaged F1-score of 0.70345, which means the performance was balanced across both "Misogyny" and "Not-Misogyny" classes. The source code for the proposed approach and found here [1].

Category	Precision	Recall	F1-score
Misogyny	0.74	0.62	0.67
Not-misogyny	0.86	0.92	0.89
Accuracy		0.84	

Table 1: Classification report for Development data

6 Limitations

The proposed method is efficient, but it has the following limitations:

- The ResNet-50 and BERT architectures are implemented without fine-tuning, which limits the model's ability to learn task-relevant features fundamental to the interpretation of misogynistic content.
- Feature fusion happens by concatenation and shallow multilayer perceptron, and does not account for complex interactions between image and text modalities.
- The classification framework provides binary output only, which does not provide the capacity to distinguish between different types of misogynistic expression or to discriminate intensity, severity, or type among misogynistic expressions.
- The model relies on provided transcriptions and does not extract text from embedded text-images to assess memes when the meme text is part of the image.

7 Conclusion

This paper described a multimodal deep learning approach to the detection of misogyny in memes by utilizing visual and textual modalities together. We took visual features from ResNet-50 and textual features from BERT. Both models used were frozen to save computational cost and training time, and we used a light-weight multilayer perceptron (MLP) to fuse the features and perform binary classification. Overall, the results exhibited that, even

[1] https://github.com/SreejaKumaravel/Misogyny-Meme-Detection

with minimal fine-tuning, the architecture was capable of capturing the implicit and explicit clues of misogyny that exist within memes. The method also proved to be very robust and simple to use for real-world deployments for harmful content moderation. Future work could include end-to-end training, different data augmentations or using cross-modal attention for a more complex fusion.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025. An overview of the misogyny meme detection shared task for chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief*, 44:108526.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *Preprint*, arXiv:2410.02378.

Sarbajeet Pattanaik, Ashok Yadav, and Vrijendra Singh. 2025. Dll5143@DravidianLangTech 2025: Majority voting-based framework for misogyny meme detection in Tamil and Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 191–199, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and 1 others. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: workshop on multimodal fact checking and hate speech detection, CEUR*.

Smriti Singh, Amritha Haridasan, and Raymond Mooney. 2023. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, Toronto, Canada. Association for Computational Linguistics.

SSNCSE@LT-EDI-2025:Speech Recognition for Vulnerable Individuals in Tamil

Sreeja K, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
sreeja2350625@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Speech recognition is a helpful tool for accessing technology and allowing people to interact with technology naturally. This is especially true for people who wants to access technology but may encounter challenges interacting with technology in traditional formats. Some examples of these people include the elderly or people from the transgender community. This research presents an Automatic Speech Recognition (ASR) system developed for Tamil-speaking elderly and transgender people who are generally underrepresented in mainstream ASR training datasets. The proposed work used the speech data shared by the task organisers of LT-EDI2025. In the proposed work used the fine tuned model of OpenAI's Whisper model with Parameter-Efficient Fine-Tuning (P-EFT) with Low-Rank Adaptation (LoRA) along with SpecAugment, and used the AdamW optimization method. The model's work led to an overall Word Error Rate (WER) of 42.3% on the untranscribed test data. A key feature of our work is that it demonstrates potential equitable and accessible ASR systems addressing the linguistic and acoustic features of vulnerable groups.

1 Introduction

Automatic Speech Recognition (ASR) has made great advancements in the form of multilingual models like OpenAI's Whisper that demonstrate solid performance across a variety of languages and acoustic conditions. Nevertheless, these systems apply in a generic sense and perform poorly when a marginalized population, such as elderly and transgender speakers, is in a low-resource language setting like Tamil. These populations may have unique vocal characteristics, vocabulary, or fluency differences that are not represented in the standard ASR model training data.

To address this disparity, we must improve current ASR systems to better understand the linguistic and phonetic diversity of the group. Recent research describes how ASR systems can introduce significant bias, especially when the target user does not fit the general speaker types used when training a model (Koenecke et al., 2020). There are even greater disparities present in languages such as Tamil because there is rarely an opportunity to directly account for acoustic variability across speaker demographics at scale.

To this end, we explore a Parameter-Efficient Fine-Tuning (PEFT) method with a pretrained version of OpenAI's Whisper model, a state-of-the-art multilingual ASR system (Radford et al., 2022). PEFT approaches like Low-Rank Adaptation (LoRA) allow efficient fine-tuning by only fine-tuning a small subset of model parameters for computational efficiency while also allowing for the performance to be maintained (Hu et al., 2021). This also makes PEFT approaches well-suited for customizing ASR systems for underrepresented user groups in low-resource settings.

We will use this technique to transcribe Tamil speech from old and transgender speakers. We wish to assess inclusive and effective options for Whisper and contribute to equitable and accessible speech technologies.

The rest of the paper is organized as follows: Section 2 analyzes the related works done in the previous research, and Section 3 discusses the speech corpus used in the current work. Section 4 contains a detailed discussion of the proposed model. Section 5 explains the experimental results. Section 6 discusses the limitations. Section 7 concludes the paper.

2 Related work

Automatic Speech Recognition (ASR) for underrepresented and marginalized communities, such as elderly individuals and the transgender community, is an area that is still difficult and underdeveloped,

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 6–10

particularly in lower-resource languages such as Tamil. (B. Bharathi, 2025) provides an overview of the Fifth shared task on Speech Recognition for Vulnerable Individuals in Tamil. In recent years, there have been increasing efforts aimed at developing ASR systems that are more inclusive for such populations – the LT-EDI shared tasks (B et al., 2022, 2023, 2024) have been very helpful in gathering and sharing annotated datasets of Tamil speech data, with elderly and transgender speakers, and setting benchmarks to evaluate models. The paper (B et al., 2025) provides an overview of this shared task on speech recognition for vulnerable people. Dynamic models such as wav2vec 2.0 and Whisper have been refined in several experiments to consider these distinct speech patterns. A study conducted by (R et al., 2024), Tamil datasets and vulnerable speakers were optimized on Whisper and XLS-R models, and reported a word error rate (WER) of 24. 45%. The wav2vec 2.0 large-xls-r300m-tamil model provided a WER of 29.30% when using speech data that included elderly speakers (Suhasini and Bharathi, 2024). Other works comparing traditional transformer-based models (BERT, RoBERTa) with wav2vec reported WERs ranging from 37.71% to 40.55% (Saranya and Bharathi, 2023). Beyond the Tamil and the LT-EDI shared tasks, international work has been done to develop ASR for underrepresented groups across languages. The authors (Hu et al., 2023) used a domain-adapted self-supervised model (beginning with), wav2vec 2.0 as a feature extractor for the TDNN and Conformer ASR systems, finding lower WERs in dysarthric speech and elderly English-speaking speech corpora. The research (Zheng et al., 2024) pursued a different approach by fine-tuning ASR systems over speech data from people with Parkinson's disease (i.e., subglottic), and enhanced models as a multitask learning architecture to produce more accurate transcriptions, as well as prediction of symptom severity. A new approach to adaptation, spectrotemporal adaptation, was used by (Geng et al., 2022), where the models were specifically fine-tuned with the speech data of dysarthric and elderly speech, which outperformed all baseline adaptation strategies. An improvement in WER for disordered speech recognition was demonstrated by (Wang et al., 2023) by hyperparameter tuning with a set of Conformer models. In low-resource contexts (e.g., elderly Frisian speakers), when the authors used data augmentation and fine-tuned wav2vec 2.0 XLS-R (and trained the models with several learning rate adjustments), the findings indicated a 20% relative WER improvement when each audio sample was 50 seconds or longer. Finally, (Chen and Asgari, 2020) used transfer learning to train ASR on socially isolated seniors and approached the problem with attention mechanisms to improve robustness given the inevitable variability (e.g., cognitive and acoustic) with this population. In general, the studies discussed in this section promoted the importance of language, domain-aware, and inclusive adaptation approaches in the construction of ASR systems for marginalized groups of people.

In this work, our approach proposed an effective Automatic Speech Recognition (ASR) system for elderly and transgender Tamil speakers by applying Parameter-Efficient Fine-Tuning (PEFT) on the Whisper model. Unlike prior work that relies on full model fine-tuning, we proposed that only a small and hopefully manageable subset of parameters is adapted, which provides a cost-effective option while maintaining similar performance to the unwieldy fine-tuning of 176 million parameters. This approach provides a good way forward as ASR can scale to lower resource settings and thus provide an avenue for developers to address inclusive ASR. Overall, this work offers a simple and low-cost pragmatic approach to providing access to communities to the speech landscape.

3 Speech Corpus Description

The dataset is comprised of spontaneous speech data collected to support Automatic Speech Recognition (ASR) for vulnerable Tamil-speaking communities (especially elders and transgender people)(B et al., 2022). People in their older years commonly attend primary points such as banks, hospitals, and administrative offices, which address their needs in their daily lives, where communication is essential. This dataset includes 7 hours and 30 minutes of spontaneous Tamil speech from people whose mother tongue is Tamil. Recordings were done in controlled environments to ensure clarity of audio quality with no background noise or overlap. All files are in .wav format. We have approximately 5.5 hours of the audio transcribed, and as the training set, 2 hours is used for testing, and is un-transcribed. Table 1 describes the data distribution for training and testing.

Data	No. of utterances	Duration in Hours
Training	908	5.5
Testing	451	2
Total	1359	7.5

Table 1: Dataset Distribution

4 Proposed Methodology

This study presents a Tamil Automatic Speech Recognition (ASR) system specifically for elderly and transgender speakers and realizes the differences in linguistic characteristics and phonetic uniqueness that exist in their spontaneous speech. The above speaker groups are noticeably underrepresented in speech data, which can lead to poor or no performance by general ASR systems. To overcome this aspect, the methodology is proposed to adapt a large, pre-trained speech model by way of a parametrically efficient methodology.

A fine-tuned version of OpenAI's Whisper model, known as yaygomii/FYP_Whisper_PEFT_TAMIL (Saranya et al., 2025), is used as the base system. This model is adapted using Low-Rank Adaptation (LoRA), a technique under the Parameter-Efficient Fine-Tuning (PEFT) framework. LoRA allows a model to learn domain-specific features by introducing trainable low-rank matrices into pre-existing attention layers while freezing all but a small number of parameters. This reduces the computational burden and memory use in the training phase, making it a great option for low-resource scenarios.

The model was fine-tuned using 5.5 hours of Tamil speech, manually transcribed from elderly and transgender speakers. Given the relatively small dataset size, this methodology incorporates data augmentation to improve model generalizability. Use of SpecAugment by using time and frequency masking of the input spectrograms is included to simulate variability in speech patterns that may be present in actual recordings from the populations of interest. Background noise is imported into the training dataset to also help improve robustness for the model to handle real environmental conditions.

The goal of these augmentation strategies was to mitigate overfitting and enhance the model's generalization to denoting expected acoustic variation in spontaneous conversation and speech. Once the model was fine-tuned, transcript files underwent post-processing to improve their linguistic and contextual accuracy. Linguistic 'ripe' corrections pertinent to Tamil used domain-specific dictionaries and language model-based rules and corrected graphemic inconsistencies, removed common spelling mistakes while keeping contextual consistency in mind for the sentence; this is very important for a morphologically rich language like Tamil.

In addition, decoder prompts can be forced during the decoding stage, so that the model remains consistent with the Tamil language, particularly in multilingual settings. The proposed strategy focuses not only on improving recognition performance for a minority group but also illustrates a commitment to the ethical and inclusive design of speech technologies.

By adjusting the system to the phonetic, lexical, and syntactic tendencies of marginalized communities, the resulting template is intended to make ASR systems more equitable and representative. The method leaves room for future improvement by proposing the use of speaker embeddings, for greater personalization and context, and the use of semi-supervised learning, to utilize the rest of the untranscribed half of the dataset. These directions continue to align with the larger purpose of developing robust and accessible ASR systems to serve many speaker populations with integrity.

5 Experimental results

The effectiveness of the proposed Tamil ASR system was evaluated, and a 2-hour test set was developed subsequently using the recordings of spontaneous elderly and transgender speech that had not been transcribed. The recordings reflected the same demographic and acoustic characteristics of the data used for training.

Word Error Rate (WER) is the main evaluation metric in this study, which is the percentage of substitutions, insertions, and deletions required to match the system's output to the true output. After decoding the output and performing post-processing with Tamil-specific corrections and dictionaries for the reference, the model was evaluated, and the WER was 42.3% over the test set. Fig. 1 demonstrates the target and predicted sentences. The source code for the proposed approach and found here [1]

[1] https://github.com/SreejaKumaravel/Speech_

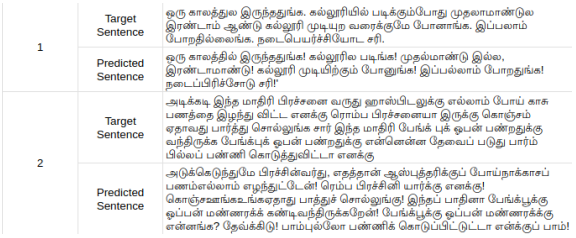

Figure 1: Sample target and predicted sentence

The model has shown evidence of transcribing Tamil speech from older adult and transgender participants while showing respect to dialectal and morphological ambiguities. Overall, the model's syntax and intention worked very well within the speech, but rare forms were difficult and therefore not transcribed well with this original achievement. The model achieved this level of performance in a resource-constrained language environment with LoRA tuning of the pre-trained model, data augmentation, and Tamil-specific post-processes.

6 Limitations

The current work faced several hurdles, including:

- Limited training makes generalization difficult across diverse communicative and acoustic varieties.
- Thus, High WER with pitying 42.3% means using rare word forms and dialectal variation, and complex sentence structures has looked difficult.
- Untranscribed data, which would prevent performance gains in semi-supervised methods.
- No speaker-specific adaptation, meaning the procedure ignores the possibility of speaker embeddings or speaker-specific personalization techniques.

7 Conclusions

This work presents an inclusive Tamil Automatic Speech Recognition (ASR) system tailored for elderly and transgender individuals, developed using a Parameter-Efficient Fine-Tuning (PEFT) approach on OpenAI's Whisper model. The system leverages a range of techniques, including data augmentation, SpecAugment, LoRA, and Tamil-specific post-processing, achieving a word error rate (WER) of 42.3% on spontaneous speech from marginalized communities. Despite limited resources, this work demonstrates the feasibility of building equitable, speech-based technologies in low-resource and underrepresented settings. Future directions include integrating speaker embeddings, employing semi-supervised learning to utilize untranscribed speech, and reducing recognition errors to further enhance accessibility for speakers with diverse speech patterns.

Recognition/blob/main/Automatic_speech_detection.ipynb

References

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya Natarajan, Rajeswari Natarajan, S Suhasini, and Swetha Valli. 2023. Overview of the second shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 31–37, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Rajalakshmi R, Suhasini S, and Swetha Valli. 2025. Overview of the fourth shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, Naples. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Sripriya N, Rajeswari Natarajan, and Suhasini S. 2024. Overview of the third shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 133–138, St. Julian's, Malta. Association for Computational Linguistics.

N. Sripriya Rajeswari Natarajan Rajalakshmi R S. Suhasini B. Bharathi, Bharathi Raja Chakravarthi. 2025. Overview of the Fifth Shared Task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Liu Chen and Meysam Asgari. 2020. Refining automatic speech recognition system for older adults. *Preprint*, arXiv:2011.08346.

Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, Shujie Hu, Xunying Liu, and Helen Meng. 2022. Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. *Preprint*, arXiv:2202.10290.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shujie Hu, Xurong Xie, Zengrui Jin, Mengzhe Geng, Yi Wang, Mingyu Cui, Jiajun Deng, Xunying Liu, and Helen Meng. 2023. Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition. *Preprint*, arXiv:2302.14564.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Jairam R, Jyothish G, Premjith B, and Viswa M. 2024. CEN_Amrita@LT-EDI 2024: A transformer based speech recognition system for vulnerable individuals in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 190–195, St. Julian's, Malta. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

S Saranya and B Bharathi. 2023. Sanbar@ lt-edi-2023: Automatic speech recognition: vulnerable old-aged and transgender people in tamil. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 155–160.

S Saranya, B Bharathi, S Gomathy Dhanya, and Aishwarya Krishnakumar. 2025. Real-time continuous tamil dialect speech recognition and summarization. *Circuits, Systems, and Signal Processing*, 44(4):2855–2881.

S Suhasini and B Bharathi. 2024. Asr tamil ssn@ lt-edi-2024: Automatic speech recognition system for elderly people. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 294–298.

Tianzi Wang, Shoukang Hu, Jiajun Deng, Zengrui Jin, Mengzhe Geng, Yi Wang, Helen Meng, and Xunying Liu. 2023. Hyper-parameter adaptation of conformer asr systems for elderly and dysarthric speech recognition. *Preprint*, arXiv:2306.15265.

Xiuwen Zheng, Bornali Phukon, and Mark Hasegawa-Johnson. 2024. Fine-tuning automatic speech recognition for people with parkinson's: An effective strategy for enhancing speech technology accessibility. In *Interspeech 2024*, interspeech$_2$024, $page$2485˘2489.$ISCA$.

CrewX@LT-EDI-2025: Transformer-Based Tamil ASR Fine-Tuning with AVMD Denoising and GRU-VAD for Enhanced Transcription Accuracy

Ganesh Sundhar S[1], Hari Krishnan N[1], Arun Prasad TD[1], Shruthikaa V[1], Jyothish Lal G[1]
[1]Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22017, cb.en.u4aie22020, cb.en.u4aie22004, cb.en.u4aie22047}
@cb.students.amrita.edu, g_jyothishlal@cb.amrita.edu

Abstract

This research presents an improved Tamil Automatic Speech Recognition (ASR) system designed to enhance accessibility for elderly and transgender populations by addressing unique language challenges. We address the challenges of Tamil ASR—including limited high-quality curated datasets, unique phonetic characteristics, and word-merging tendencies—through a comprehensive pipeline. Our methodology integrates Adaptive Variational Mode Decomposition (AVMD) for selective noise reduction based on signal characteristics, Silero Voice Activity Detection (VAD) with GRU architecture to eliminate non-speech segments, and fine-tuning of OpenAI's Whisper model optimized for Tamil transcription. The system employs beam search decoding during inference to further improve accuracy. Our approach achieved state-of-the-art performance with a Word Error Rate (WER) of 31.9,winning first place in the LT-EDI 2025 shared task.

Keywords: Speech Recognition, Transformer, Whisper, Adaptive Variational Mode Decomposition, Vulnerable populations, Low-resource, Dravidian language.

1 Introduction

Communication has been the cornerstone of human evolution, enabling individuals to share thoughts, emotions, and information. Human communication evolved naturally through verbal expression, with speech emerging as one of our earliest and most intuitive forms of interaction. With recent advancements in technology, a more natural way for human-computer interaction is necessary, which is satisfied with the help of speech processing techniques. This has led to the invention of Automatic Speech Recognition (ASR) (Yu and Deng, 2016) systems. The goal of an ASR system is to convert spoken language into text and it plays a crucial role in various applications such as virtual assistants, transcription services, and accessibility tools.

Despite advancements in technology, ASR still remains a complex task due to several challenges. The high variability in speech data makes it difficult for the models to generalize across different speakers. Additionally, while ASR systems for languages like English, Spanish and Mandarin benefit from extensive datasets and pretrained models, Dravidian languages like Tamil and Malayalam suffer from the scarcity of extensive & good quality annotated datasets, making it hard to build robust ASR systems. Tamil, in particular, has unique phonetic characteristics and word-merging tendencies that further complicate transcription accuracy (Akhilesh et al., 2022; Shraddha et al., 2022).

This work addresses these challenges by fine-tuning a Transformer-based ASR system for Tamil speech, specifically for vulnerable old-aged and transgender individuals, who often face difficulties in accessing essential services due to lack of literacy and familiarity with technology. To enhance transcription accuracy, our method integrates Adaptive Variational Mode Decomposition (AVMD) (Lian et al., 2018) for noise reduction and Silero Voice Activity Detection (VAD) (Team, 2021) to eliminate non-speech segments. We then, finetune OpenAI's Whisper (Radford et al., 2023) model for Automatic Speech Recognition, leading to promising results. This approach achieved state-of-the-art (SOTA) performance, ranking first in the shared task with a Word Error Rate (WER) of 31.9. This demonstrates the effectiveness of this pipeline in handling the complexities of Tamil speech for vulnerable populations (Chowdary et al., 2024).

2 Related Work

Early ASR systems relied on rule-based phonetic models and statistical methods like Hidden Markov Models (HMMs) (Levinson, 1986), Gaussian Mix-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 11–16

ture Models (GMMs) (Gorin et al., 2014) and Dynamic Time Warping (DTW) (Nair and Sreenivas, 2008). However, these methods struggled with high variability in speech, such as speaker, accents, age, gender, background noise, and spontaneous speech patterns. This led to the emergence of traditional deep learning based models such as Recurrent Neural Networks (RNNs) (Jain et al., 2020) and Long Short-Term Memory (LSTM) (Weninger et al., 2015) networks. (Dahl et al., 2011) used Deep Neural Network and HMM based hybrid model which further improved the quality of the transcriptions generated.

But even they struggled with capturing long range dependencies. The major breakthrough in processing temporal was made with the invention of the Transformer architecture (Vaswani et al., 2017) in 2017. It introduced the Multi Head Self Attention (MHSA) mechanism which excelled in capturing both long term and short term dependencies. Hence, the Transformer-based models significantly improved transcription accuracy by learning complex patterns from large datasets. Self supervised models like Wav2Vec 2.0 (Baevski et al., 2020) reduced reliance on labeled data by learning speech representations from raw audio, significantly improving ASR for low-resource languages. OpenAI's Whisper further advanced ASR with large-scale weak supervision, enabling robust transcription, translation, and language identification across diverse datasets (Barathi Ganesh et al., 2024).

3 Dataset

Elderly individuals frequently visit essential service locations such as banks, hospitals, and administrative offices but struggle with technological tools designed to assist them. Similarly, transgender individuals often face educational barriers due to societal prejudices, making speech a primary mode of communication for accessing essential services. Hence, the dataset provided for this shared task specifically targets vulnerable elderly and transgender individuals (B et al., 2022). The dataset consists of 7.5 hours of spontaneous Tamil speech, which is split into 5.5 hours of transcribed speech for training and 2 hours of unlabeled speech for testing. The train data provided was further made into a 80-20 split for training and development. The dataset distribution is provided in Table 1

Split	Audios	Duration (hours)
Train	726	4.4
Dev	182	1.1
Test	451	2.0
Total	1,359	7.5

Table 1: Dataset distribution of Tamil ASR corpus

4 Methodology

The speech signal was initially denoised using Adaptive Variational Mode Decomposition (AVMD), which decomposes it into variational modes and reconstructs the relevant components to remove noise while preserving speech clarity. Next, Silero VAD (GRU based) was applied to eliminate silence and non-speech segments, reducing computational load due to unnecessary processing and at the same time improving transcription accuracy.The processed audio is then passed through the Whisper processor, which converts it into log-Mel spectrogram (Stevens et al., 1937) features using a standardized pipeline. Finally, the extracted features are passed to the Whisper-Tamil-Medium model for generating transcriptions, with beam search decoding (Lowerre, 1976) during testing to enhance the accuracy and reduce the WER. The overall workflow is depicted in Figure 1.

4.1 AVMD - Denoising

The dataset contains multiple audio samples, where some are of high quality while others suffer from a bit of background noise, which can affect the transcription process. Specifically, audio files in Series 3 (*Audio-3_1.wav* to *Audio-3_32.wav*) exhibit excellent clarity, whereas Series 2 (*Audio-2_1.wav* to *Audio-2_26.wav*) contains noticeable background noise. To account for this, Adaptive Variational Mode Decomposition (AVMD) was selectively applied based on a noise parameter threshold (Signal to Noise Ratio (Johnson, 2006)) , ensuring only noisy signals underwent processing. Unlike traditional Variational Mode Decomposition (VMD) (Dragomiretskiy and Zosso, 2013), which uses fixed decomposition parameters, AVMD dynamically adjusts mode selection and bandwidth constraints based on the signal's characteristics, making it more effective in separating noise components from speech. This method outperforms other denoising methods such as wavelet denoising (Luo et al., 2012) or spectral subtraction (Martin, 1994), which often introduce artifacts.

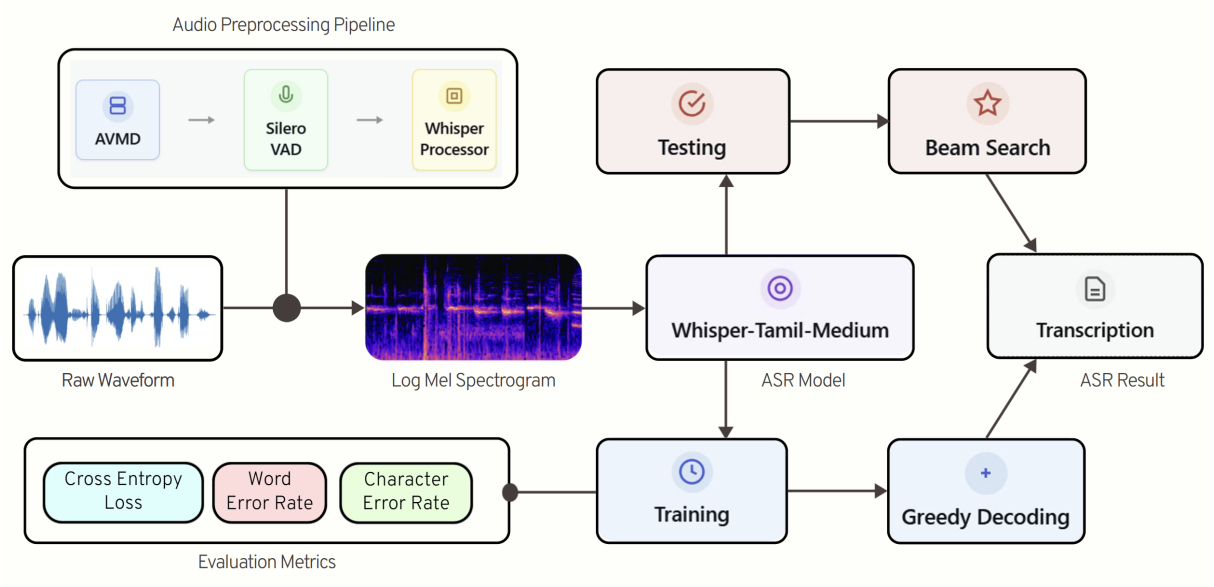

Figure 1: End to end pipeline for Tamil ASR using the Whisper-Tamil-Medium model. The raw audio is preprocessed using AVMD, Silero VAD, and the Whisper Processor, transforming it into Log Mel Spectrograms. These spectrograms serve as input to the Whisper model, which undergoes training and evaluation with greedy decoding and beam search techniques. The resulting transcriptions are assessed using cross-entropy loss, WER, and CER to measure performance.

4.2 Silero VAD

Some of the audio samples in the dataset contained long pauses, which negatively impacted ASR performance. So, any segment with silence exceeding 300 ms was removed, such as in *Audio-1_01.wav*, where a pause from 12s to 14s was eliminated. Silero VAD was chosen for this task due to its robust deep learning based Voice Activity Detection (VAD), outperforming traditional energy-based or statistical methods (Sohn et al., 1999). The implementation involved loading the pretrained Silero VAD model, detecting speech timestamps, merging overlapping or close speech segments, and extracting speech regions while preserving 300 ms of buffer before and after detected speech to compensate for potential mis-detections where speech might be partially cropped.

4.3 Whisper Finetuning

We chose the pretrained hugging face Whisper model from *"vasista22/whisper-tamil-medium"* as it achieved the SOTA performance even without preprocessing in the same shared task last year (Jairam et al., 2024). The selected model was pretrained on various tamil ASR datasets such as *IISc-MILE Tamil ASR Corpus* (A et al., 2022), *ULCA ASR Corpus*, etc. Whisper is a transformer-based ASR model that processes audio using a feature extractor that converts raw waveforms into mel spectrograms, capturing key frequency components over time. The transcriptions given were converted to tokens with the help of the Whisper Tokenizer. The model begins with a CNN-based feature extractor (O'shea and Nash, 2015) layer, where two 1D convolution layers (Conv1d) expand the input from 80 mel filter banks to 1024 channels, followed by downsampling. The processed audio features are directly fed to the Whisper encoder then passed to the decoder, both containing 24 Transformer layers for transcription generation. The decoder fetches the features from the encoder with the help of cross-attention with the encoder memory. The final linear layer maps the decoder output to the vocabulary space, generating transcriptions (B et al., 2025).

4.4 Beam Search Decoding

Beam search decoding is an exploratory search technique employed in ASR systems to identify the optimal output sequence. In contrast to greedy decoding, which chooses the maximum probability token per step, beam search tracks several potential sequences (beams) during each decoding phase, minimizing errors resulting from choices that appear optimal locally but prove suboptimal overall. It is used only during test time to refine the output by considering multiple possible candidate states. During training, it is not required since the model is optimized using teacher forcing, where the correct

target sequence is provided at each step. Training prioritizes computational efficiency, while beam search during inference ensures higher-quality predictions.

5 Experimentation

For training, the model was optimized using *AdamW* (Loshchilov and Hutter, 2017) with a learning rate of 10^{-5} and weight decay of 10^{-2} to prevent overfitting. The loss function used was *CrossEntropyLoss*, which measures the difference between predicted and actual token distributions. We employed ReduceLROnPlateau to automatically decrease the learning rate by half whenever validation loss stopped improving for two consecutive epochs, which helped maintain smooth training progress. The models were trained on a *4080 Super* GPU with 4 as the batch size. During testing, beam search decoding was employed, thus increasing transcription accuracy. The trends observed during the training and validation process are given in Figure 2.

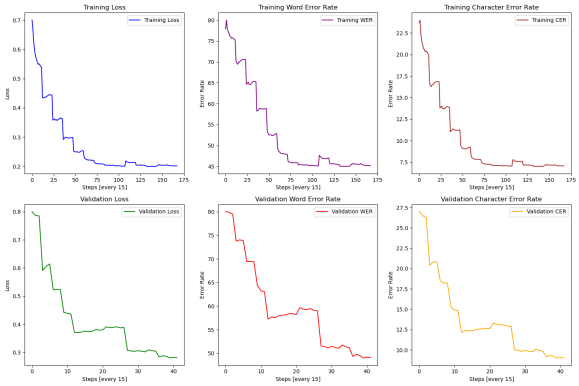

Figure 2: Training and validation performance metrics for the best-performing Tamil ASR model.

6 Result and Analysis

6.1 Word Error Rate

To evaluate the ASR pipeline built, Word Error Rate (WER) (Levenshtein et al., 1966) was used, as it is the metric specified for comparing the results in the shared task. It quantifies the transcription accuracy by comparing the words in the system's output with the reference transcript. The WER is defined as

$$WER = \frac{S + D + I}{N} \quad (1)$$

where S represents the number of substitutions (incorrectly transcribed words), D denotes deletions (omitted words), I accounts for insertions (extra words added), and N is the total number of words in the original reference transcript. A lower value of WER generally indicates better performance.

6.2 Model Evaluation

The pipeline built was assessed using three different metrics: Cross Entropy Loss, Word Error Rate (WER), and Character Error Rate (CER) (Rice et al., 1995). The results of the best performing model in terms of WER are summarized in Table 2.

Metric	Train	Validation	Test
Loss	0.202	0.281	-
WER	45.212	49.095	31.9
CER	7.061	9.037	-

Table 2: Performance Metrics of the Best Model

6.3 Result Comparison

Our proposed methodology, integrating Adaptive Variational Mode Decomposition (AVMD) for denoising, Silero VAD for voice activity detection, Whisper for transcription, and Beam Search decoding, achieved a WER of 31.9 on the testing set, securing the 1st rank in the competition. The scores of the top five performing teams in the shared task are summarized in Table 3.

Rank	Team Name	WER (%)
1	CrewX	31.90
2	NSR	34.85
3	Wictory	34.93
4	JUNLP	38.42
5	SSNCSE	42.30

Table 3: Top 5 Teams scored based on Word Error Rate

7 Conclusion

In this paper, we presented a robust ASR pipeline tailored for noisy audio conditions. In conclusion, our methodology showcases the potential of combining noise-aware preprocessing and advanced decoding strategies to deliver accurate and reliable transcription in real-world, low-resource scenarios. This also provides a strong foundation for future enhancements in speech recognition under challenging conditions. **The entire implementation can be found here:** https://github.com/Ganesh2609/VulnerableSpeechASR

8 Limitations

Even though the proposed pipeline performed well in terms of the WER, it had a few noticeable limitations, which are as follows:

1. Some audio samples had severely distorted speaker voices, making them unintelligible even to human listeners. As a result, the model also struggled to transcribe such cases accurately.
2. The model occasionally merges separate Tamil words into a single word. For example, *vandhu irunthaal* is sometimes transcribed as *vanthirunthaal*, which may lead to a lesser WER.
3. The training dataset was relatively small for ASR tasks, with 908 audio samples for training and 451 for testing (approximately a 66.7-33.3 split). Creating a development set from the training data reduced the effective training size further, leading to signs of overfitting during training.

References

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. 2022. Subword dictionary learning and segmentation techniques for automatic speech recognition in tamil and kannada. *arXiv preprint*.

A Akhilesh, P Brinda, S Keerthana, Deepa Gupta, and Susmitha Vekkot. 2022. Tamil speech recognition using xlsr wav2vec2. 0 & ctc algorithm. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Rajalakshmi R, Suhasini S, and Swetha Valli. 2025. Overview of the fourth shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, Naples. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

HB Barathi Ganesh, G Jyothish Lal, R Jairam, KP Soman, NS Kamal, and B Sharmila. 2024. Corepool—corpus for resource-poor languages: Badaga speech corpus. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 193–211.

Divi Eswar Chowdary, Rahul Ganesan, Harsha Dabbara, G Jyothish Lal, and B Premjith. 2024. Transformer-based multilingual automatic speech recognition (asr) model for dravidian languages. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 259–273.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2011. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42.

Konstantin Dragomiretskiy and Dominique Zosso. 2013. Variational mode decomposition. *IEEE transactions on signal processing*, 62(3):531–544.

Arseniy Gorin, Denis Jouvet, Emmanuel Vincent, and Dung Tran. 2014. Investigating stranded gmm for improving automatic speech recognition. In *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pages 192–196. IEEE.

Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf. 2020. Contextual rnn-t for open domain asr. *arXiv preprint arXiv:2006.03411*.

R Jairam, G Jyothish, B Premjith, and M Viswa. 2024. Cen_amrita@ lt-edi 2024: A transformer based speech recognition system for vulnerable individuals in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 190–195.

Don H Johnson. 2006. Signal-to-noise ratio. *Scholarpedia*, 1(12):2088.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Stephen E Levinson. 1986. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech & Language*, 1(1):29–45.

Jijian Lian, Zhuo Liu, Haijun Wang, and Xiaofeng Dong. 2018. Adaptive variational mode decomposition method for signal processing based on mode characteristic. *Mechanical Systems and Signal Processing*, 107:53–77.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Bruce T Lowerre. 1976. *The harpy speech recognition system.* Carnegie Mellon University.

Guomin Luo, Daming Zhang, and DD Baleanu. 2012. Wavelet denoising. *Advances in wavelet theory and their applications in engineering, physics and technology*, pages 59–80.

Rainer Martin. 1994. Spectral subtraction based on minimum statistics. *power*, 6(8):1182–1185.

Nishanth Ulhas Nair and TV Sreenivas. 2008. Multi pattern dynamic time warping for automatic speech recognition. In *TENCON 2008-2008 IEEE Region 10 Conference*, pages 1–6. IEEE.

Keiron O'shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Stephen V Rice, Frank R Jenkins, and Thomas A Nartker. 1995. The fourth annual test of ocr accuracy. Technical report, Technical Report 95.

S Shraddha, Sachin Kumar, et al. 2022. Child speech recognition on end-to-end neural asr models. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6. IEEE.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3.

Stanley Smith Stevens, John Volkmann, and Edwin B Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. *Retrieved March*, 31:2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. 2015. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*, pages 91–99. Springer.

Dong Yu and Lin Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

JUNLP@LT-EDI-2025: Efficient Low-Rank Adaptation of Whisper for Inclusive Tamil Speech Recognition Targeting Vulnerable Populations

Priyobroto Acharya[1], **Soham Chaudhuri**[2], **Sayan Das**[3],**Dipanjan Saha**[4], **Dipankar Das**[5]
[1]Dept. of Power Engineering, Jadavpur University, Kolkata, India
[2]Dept. of Electrical Engineering, Jadavpur University, Kolkata, India
[3,4,5]Dept. of CSE, Jadavpur University, Kolkata, India
{ priyobrotoacharya98, sohamchaudhuri.12.a.38,sayan.das200216,sahadipanjan6, dipankar.dipnil2005}@gmail.com

Abstract

Speech recognition has received extensive research attention in recent years. It becomes much more challenging when the speaker's age, gender and other factors introduce variations in the speech. In this work, we propose a fine-tuned automatic speech recognition model derived from OpenAI's whisper-large-v2. Though we experimented with both Whisper-large and Wav2vec2-XLSR-large, the reduced WER of whisper-large proved to be a superior model. We secured **4**th rank in the LT-EDI-2025 shared task. Our implementation details and code are available at our **GitHub repository**[1].

1 Introduction

Automatic Speech Recognition (ASR) has transformed the way humans interact with machines by enabling devices to understand spoken language. It plays a crucial role in enhancing accessibility for individuals with disabilities, such as the elderly and those with hearing or speech impairments(Yu and Deng, 2017; Malik et al., 2021). By allowing voice-based interaction, ASR improves ease of communication and overall quality of life for these groups.

While ASR systems have achieved impressive accuracy in languages like English, low-resource languages such as Tamil still face challenges (Ramesh and Gupta, 2021). Tamil, spoken by millions across Tamil Nadu, Sri Lanka, and Singapore, is linguistically rich and features numerous regional dialects, making speech recognition particularly complex. These challenges are amplified when recognizing speech from vulnerable populations, such as those with dysarthria or slurring (Christensen, 2013).

In this work, we focus on building an inclusive Tamil ASR system by fine-tuning the Whisper model (vasista22/whisper-tamil-large-v2), known for its strong multilingual performance (Radford et al., 2022). To make the fine-tuning process efficient, we use Low-Rank Adaptation (LoRA), which reduces the computational burden while maintaining high accuracy (Hu et al., 2021). Our training dataset includes Tamil speech samples from diverse dialects and speakers with impairments. The fine-tuned model achieves a Word Error Rate (WER) of **38.42%**, demonstrating significant improvement and the potential of Whisper models in developing accessible ASR systems for underrepresented languages.

2 Related Work

Automatic speech recognition (ASR) has evolved from hybrid Hidden Markov Model-Gaussian Mixture Model (HMM-GMM)(Xuan et al., 2001) frameworks to end-to-end deep learning systems. Early systems leveraged HMMs for temporal modeling and DNNs for acoustic feature extraction, achieving significant accuracy improvements over traditional methods. Transitioning to architectures like LSTMs and transformers enabled better sequential context capture, with models like Conformer and ContextNet integrating convolutional and self-attention mechanisms for spectral and global dependencies(Prabhavalkar et al., 2021). Self-supervised learning paradigms, such as wav2vec 2.0, further advanced low-resource ASR by leveraging unlabeled data for robust feature learning(Mainzinger and Levow, 2024)(Kheddara et al., 2024).

Recent efforts focus on domain-specific challenges, including elderly and vulnerable populations as well as low-resource speech recognition. (B et al., 2022)(Bartelds et al., 2023)presented findings from a shared task on Tamil ASR for vulnerable individuals, emphasizing the difficulty of recognizing atypical speech patterns in elderly

[1]https://github.com/Priyobroto98/ASR-Tamil-LTEDI-2025

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 17–25

and impaired speakers. Their work demonstrated the utility of HMM-DNN hybrid systems(Wang et al., 2019) and end-to-end models alongside data augmentation and transfer learning to improve robustness. In a follow-up shared task, (B et al., 2025) expanded the dataset and evaluated multilingual models (e.g., XLS-R, Whisper), showing that fine-tuning, domain adaptation, and acoustic normalization techniques effectively addressed speech variations and noise in low-resource settings. Similar advances include acoustic model adaptation using age-specific corpora like EARS and VOTE400, which reduce word error rates (WER) by 25% for elderly speech by mitigating spectral and prosodic variations. For low-resource languages, techniques like self-training and text-to-speech augmentation improve WER by 20–25%, as demonstrated for Gronings and Mvskoke. Transformer-based streaming architectures, employing time-restricted attention, balance latency and accuracy, while hybrid HMM-DNN systems remain relevant for stable frame-level processing. Despite progress, challenges persist in dataset diversity, real-time adaptation, and computational efficiency for edge deployment.

3 Dataset Description and Analysis

The dataset focuses on addressing the challenges faced by vulnerable groups, specifically elderly individuals and transgender people in Tamil-speaking communities, where elderly individuals often encounter difficulties using digital tools in essential locations like banks, hospitals, and administrative offices, where speech-based systems could significantly ease their interactions (Gales et al., 2019; Liu and Lutters, 2021). Similarly, transgender individuals, frequently deprived of primary education due to societal prejudice, rely heavily on speech as their primary mode of communication (Pandey and Mishra, 2019; Bose et al., 2019). By capturing the spontaneous speech patterns of these groups, the dataset aims to facilitate the development of inclusive and accessible ASR systems that cater to their unique linguistic needs and daily life challenges (Albanie et al., 2020; Srinivasan et al., 2023).

The dataset contains **908 samples** totaling nearly 5 hours of speech. We have split the entire corpus into training (894 samples, 4.87 hours), validation (9 samples, 0.05 hours), and test sets (5 samples, 0.03 hours) for tracking the performance metrics at different stages of model development. In addition to this, we were provided with **2 hours** of high-quality audio speech data, which will be used for testing purposes after successfully training our best model and following best practices.

Set	Samples	Duration (hours)	Avg Duration (seconds)	Avg Text Length (chars)
Training	894	4.87	19.61	212
Validation	9	0.05	20.00	256
Test	5	0.03	20.00	229

Table 1: Dataset Statistics and Composition

We conduct **spectrogram analysis(Khodzhaev, 2024)** on the speech dataset to characterize the time-varying frequency properties of the audio signals. In figure-1 the analysis confirms that all samples exhibit dominant speech energy below **4 kHz**, with clearly observable formant structures.

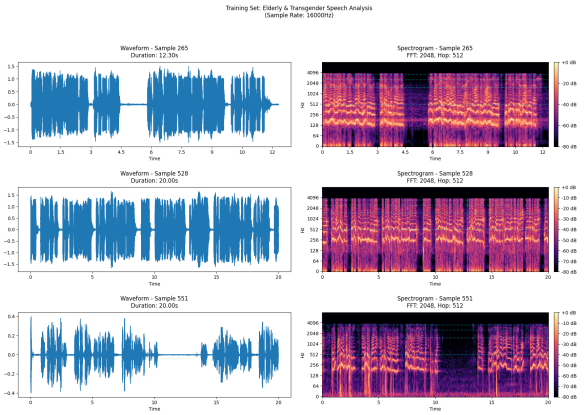

Figure 1: Representative spectrograms illustrating dominant speech energy and formant structures.

The overall spectral clarity and low background noise across all samples suggest high-quality recordings. These observations not only confirm the suitability of the data for further speech processing tasks—such as automatic speech recognition or speaker profiling (Nagrani et al., 2017; Yu et al., 2021), but also highlight the diversity in speaking styles and potential demographic differences among the speakers (Narayanan and Georgiou; **?**). Such variability is crucial for developing robust and inclusive speech systems that generalize well across different populations.

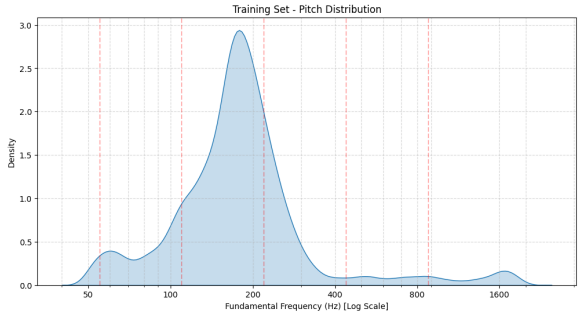

Figure 2: Pitch Distribution

In figure-2 the pitch distribution(Deruty et al., 2025) graph reveals a clear multimodal pattern, with a **dominant peak** near **200 Hz** and **secondary peaks** around **100 Hz** and at higher frequencies, indicating demographic diversity. The use of a logarithmic x-axis reflects the perceptual nature of pitch. Variations in peak heights highlight gender imbalance, which may introduce bias in ASR performance toward dominant voice types.

Figure 3: Silence Percentage Distribution

The dataset exhibits a bell-shaped silence distribution (jin Shim et al., 2024) (**mean 27.6%, median 26.4%**) with a right skew, where most samples contain **10–50%** silence (peaking at 25–30%) under a 30 dB/16 kHz detection threshold (refer Figure 3). This aligns with natural speech patterns, where pauses constitute approximately one-quarter of spoken content (Gold and Morgan, 2000), informing ASR design for effective endpoint detection and robustness (Ramírez et al., 2007). The balanced silence distribution facilitates training on realistic speech rhythms and timing structures (Jurafsky and Martin, 2000), improving temporal generalization in deployment scenarios.

From the analysis of temporal features (Figure 4), we found the audio dataset exhibits high-quality temporal features with segmented speech (amplitude ± 1.5 units) and precise silence intervals, evidenced by RMS energy drops to zero and spectral rolloff between 500–3500 Hz. Stable

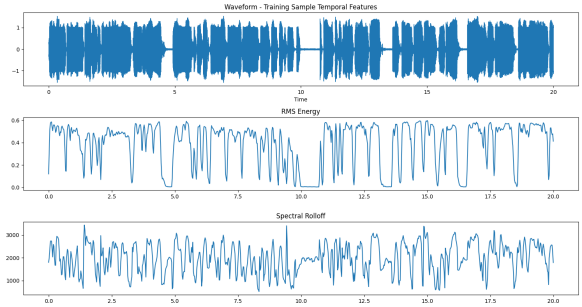

Figure 4: Audio Training Sample Temporal Features

RMS levels ($\sim$**0.4–0.5**) during speech segments indicate consistent articulation, while rolloff variations (**1000–3000 Hz**) reflect phonetic diversity, demonstrating complementary temporal-spectral features (waveform, energy, rolloff) that reveal controlled recording conditions ideal for training robust speech models requiring precise acoustic characterization (Rabiner and Schafer, 1978; Tolonen and Karjalainen, 2000; Purwins et al., 2019; Zhang et al., 2021).

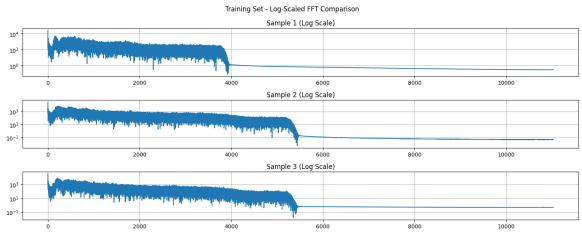

Figure 5: Log-Scaled FFT Comparison in Training Dataset

The **log-scaled FFT analysis** of the training dataset reveals concentrated spectral energy (**10^1–10^4** magnitude) in lower frequencies (**0–4000 bins**) with a sharp roll-off at **4000–5000 bins** across samples, indicating bandwidth-limited audio rich in harmonic content (refer Figure 5). Consistent noise floors (10^{-1}–10^0 magnitude) and spectral homogeneity suggest uniform recording/post-processing conditions, while the preserved harmonic structures and logarithmic energy distribution (aligning with auditory perception) highlight key perceptual features of speech signals (Choi et al., 2018; Deller et al., 1993; Verhelst and Roelands, 2000; Purwins et al., 2019).

4 Methodology and Implementation Details

In this study, speech recognition was performed using two pre-trained state-of-the-art models, Whis-

per and XLSR. Both models were trained on the Tamil corpus, and the best results were submitted for the competition.

The Whisper model (Radford et al., 2023) is a pre-trained automatic speech recognition (ASR) model trained on **680,000 hours** of multilingual and multitask supervised data sourced from the web. In our work, we have utilized **`vasista22/whisper-tamil-large-v2`**[2], which is a fine-tuned version of **`openai/whisper-large-v2`**[3] on the Tamil data available from multiple publicly available ASR corpora. This transformer-based encoder-decoder model processes log-Mel spectrograms through convolutional layers in the encoder and generates text autoregressively in the decoder. The model was further fine-tuned on a Tamil corpus of the given training dataset, providing a robust baseline for Tamil speech recognition.

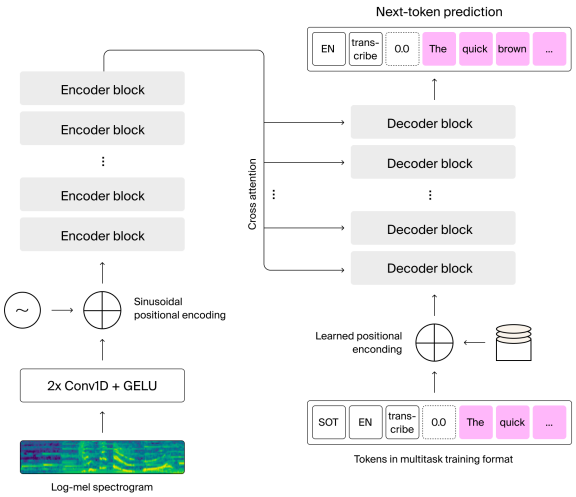

Figure 6: Whisper Model Architecture (https://openai.com/index/whisper/)

To adapt the 1.59-billion-parameter Whisper model efficiently, we utilize ***Low-Rank Adaptation (LoRA) (Hu et al., 2021)*** and ***Dynamic Rank Adaptation (DoRA) (Liu et al., 2024)***. These techniques freeze pre-trained weights and inject trainable low-rank matrices into specific transformer submodules, reducing computational overhead while preserving model performance (Xu et al., 2023).

LoRA decomposes weight updates (ΔW) into two low-rank matrices $\mathbf{A}$ and $\mathbf{B}$, where $\Delta W = \mathbf{BA}$. For a weight matrix $W \in R^{d\times k}$, the adapted weights become:

$$W' = W + \Delta W$$
$$= W + \mathbf{B}\cdot\mathbf{A}, \quad \mathbf{B} \in R^{d\times r}, \quad \mathbf{A} \in R^{r\times k}$$

where $r \ll \min(d,k)$ is the rank of adaptation. This reduces trainable parameters from $\mathcal{O}(dk)$ to $\mathcal{O}(r(d+k))$.

We apply LoRA to the query, key, value, and output projection layers of each transformer block. To ensure stable training, weight scaling is used:

$$\Delta W = \alpha \cdot \frac{\mathbf{BA}}{r} \tag{1}$$

where α is a scaling factor (typically $\alpha \in [8, 32]$), introduced to stabilize updates for small r.

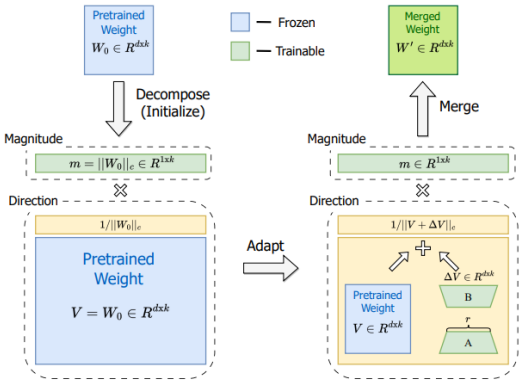

Figure 7: An overview of our proposed DoRA, which decomposes the pre-trained weight into magnitude and direction components for fine-tuning, especially with LoRA to efficiently update the direction component. Note that $\|\cdot\|_c$ denotes the vector-wise norm of a matrix across each column vector

DoRA extends LoRA by dynamically adjusting the rank r during training (Liu et al., 2024). It decomposes weights into magnitude (m) and direction ($\mathbf{V}$) components:

$$W = m \cdot \frac{\mathbf{V}}{\|\mathbf{V}\|_F} \tag{2}$$

where $\|\mathbf{V}\|_F$ is the Frobenius norm. During backpropagation, the gradient flows primarily through the direction $\mathbf{V}$, enabling more expressive parameterization even at low ranks.

Quantization to 8-bit precision was implemented using:

$$\mathbf{W}_{\text{int8}} = \text{quantize}\left(\frac{\mathbf{W} - \mu_{\mathbf{W}}}{\sigma_{\mathbf{W}}}\right)$$

where:

[2]https://huggingface.co/vasista22/whisper-tamil-large-v2
[3]https://huggingface.co/openai/whisper-large-v2

- $\mathbf{W}$ is the original full-precision weight matrix or tensor.
- $\mu_{\mathbf{W}}$ is the mean of the weight tensor $\mathbf{W}$, used for centering.
- $\sigma_{\mathbf{W}}$ is the standard deviation or scale factor of $\mathbf{W}$, used for normalization.
- $\mathbf{W}_{\text{int8}}$ is the quantized 8-bit integer representation of the normalized weights.
- $\hat{\mathbf{W}}$ is the dequantized approximation of the original weights in floating point.
- quantize$(\cdot)$ maps a real-valued input to discrete 8-bit integer levels (usually in the range $[-128, 127]$).

followed by dequantization:

$$\hat{\mathbf{W}} = \sigma_{\mathbf{W}} \cdot \mathbf{W}_{\text{int8}} + \mu_{\mathbf{W}}$$

Training employed mixed-precision arithmetic (FP16) with the **AdamW** optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$), a learning rate of 10^{-5} with 50 warmup steps, and gradient accumulation over 2 steps. Only **2.99%** of parameters (47.5M out of 1.59B) were trainable through selective application of LoRA to the query, key, and value projection layers.

During implementation, a comprehensive data preprocessing pipeline was constructed using WhisperProcessor components, which extract audio features with a sampling rate of **16kHz** and prepare corresponding text transcriptions for supervised training. We have used a custom `DataCollatorSpeechSeq2SeqWithPadding` that effectively handles variable-length audio inputs and properly masks padding tokens in labels with -100 to be ignored during loss calculation. The combined use of 8-bit quantization, LoRA, and DoRA reduced memory requirements by 4 times compared to full-precision fine-tuning and achieved a **97%** reduction in trainable parameters without significant accuracy degradation, demonstrating the efficacy of parameter-efficient methods (Dettmers et al., 2023) for large-scale ASR (Radford et al., 2023) adaptation.

On the other hand, we fine-tuned the pretrained **anuragshas/wav2vec2-xlsr-53-tamil**[4] checkpoint with the Hugging Face `Trainer` API. The model is

[4]https://huggingface.co/anuragshas/wav2vec2-xlsr-53-tamil

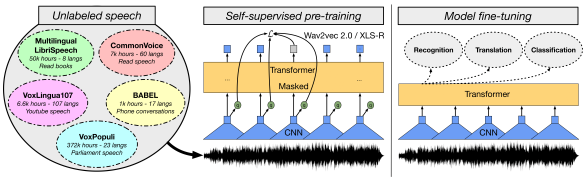

Figure 8: Fine-tuning XLSR for Tamil ASR with Transformers. (https://huggingface.co/blog/fine-tune-xlsr-wav2vec2)

a `Wav2Vec2ForCTC` type model (Conneau et al., 2021) and was fine-tuned with full-scale fine-tuning, without layer freezing or modifications. Connectionist Temporal Classification (CTC) loss was used during training and performance was tracked with Word Error Rate (WER) and Character Error Rate (CER). Mixed precision training was activated with `fp16=true`, and the best model was chosen based on the minimum WER on the evaluation set. Gradient accumulation with an accumulation step of 2 was used to stabilize training and mimic larger batch sizes.

5 Result and Discussion

Submissions to the Shared Task on Speech Recognition for Vulnerable Individuals in Tamil were evaluated using the **Word Error Rate (WER)** between the ASR hypotheses and the reference human transcriptions for the evaluation set (Morris et al., 2004).

$$\text{WER} = \frac{S + D + I}{N}$$

Where: S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference transcriptions.

During the fine-tuning phase, a close watch was kept on the WER and **Character Error Rate (CER)** of both models, which were trained for the same number of epochs (Hori et al., 2017).

Model	Val. Loss	WER(%)	CER(%)
whisper-tamil-large-v2	0.540	69.4	26.1
wav2vec2-large-xlsr-53-tamil	1.727	94.0	44.2

Table 2: ASR Model Performance Comparison

We compared both the models' WER and CER. Since the `whisper-tamil-large-v2` model

demonstrated significantly lower WER and CER than the `wav2vec2-large-xlsr-53-tamil` model, we selected it for generating transcriptions for the test dataset and submitted those results for final evaluation.

Team Name	WER	Rank
CrewX	31.9	1
NSR	34.85	2
Victory	34.93	3
JUNLP	38.42	4
SSNCSE	42.3	5

Table 3: Team-wise WER and Rank

We achieved a WER of **38.42** on the test dataset, which helped us secure the **4th** rank in the shared task. This performance demonstrates the robustness of parameter-efficient fine-tuning strategies for multilingual ASR tasks on low-resource and demographically sensitive datasets (Hsu et al., 2021).

6 Limitations

Despite its contributions, this work has several limitations. The dataset's limited size and dialectal diversity may hinder generalization, particularly for underrepresented Tamil accents (Addanki et al., 2022). Computational constraints restricted the exploration of more complex architectures and large-scale training (Gaido et al., 2021). Evaluation primarily relied on WER, which may not fully reflect real-world intelligibility or user-centric performance, especially for vulnerable populations (Falk and Chan, 2007; Meng et al., 2021). The model's performance varied across regional pronunciations, suggesting a need for more balanced data. Additionally, the absence of human-centered evaluations, such as user studies or error analysis on critical phrases, limits insights into practical usability (Amershi et al., 2019). Resource limitations also prevented extensive hyperparameter tuning and ablation studies. Broader metrics, including semantic accuracy and user satisfaction, could better assess assistive utility (Baker et al., 2020). Finally, ethical considerations, such as bias mitigation and inclusivity in data collection, were not thoroughly examined (Hovy and Prabhumoye, 2021). Addressing these gaps in future work could enhance robustness and fairness in Tamil speech recognition.

7 Future Scope

To overcome these limitations and extend the impact of this study, several avenues for future work are proposed. Expanding the dataset to include speakers from a wide range of demographics and regions, as well as recording audio in diverse environmental conditions, could enhance the model's robustness and adaptability (Ko et al., 2017; Besacier et al., 2014). Incorporating advanced architectures and exploring multilingual frameworks may further improve performance (Pratap et al., 2020; Conneau et al., 2021). Real-world deployment possibilities, such as live transcription services and language learning tools for vulnerable groups, offer practical applications of this research (Albanie et al., 2020; Srinivasan et al., 2023). Collaborations with local communities and organizations to co-develop datasets and validate findings can ensure inclusivity and greater acceptance of the model in real-world scenarios (Bender et al., 2021).

8 Conclusion

This work presents JUNLP's efficient approach to building an inclusive Tamil Automatic Speech Recognition (ASR) system for vulnerable populations, including elderly and transgender speakers. Using parameter-efficient fine-tuning (PEFT) methods Low-Rank Adaptation (LoRA) and Dynamic Rank Adaptation (DoRA), we adapted the multilingual Whisper-large-v2 model for low-resource Tamil speech with demographic variation. Our model achieved a Word Error Rate (WER) of 38.42% on the LT-EDI-2025 evaluation set, securing 4th place. By freezing Whisper's 1.59B pretrained weights and injecting low-rank matrices, we reduced trainable parameters by 97% (47.5M) and memory usage by 4 times, enabling fine-tuning on limited hardware. DoRA's decomposition improved expressiveness, and 8-bit quantization with mixed-precision training stabilized optimization. Trained on 908 speech samples (5 hours) reflecting dialectal diversity, the model showed promise in inclusive ASR. Limitations include dataset size, regional bias, and reliance on WER. Future directions include expanding diverse corpora and integrating user-centered evaluations. This study affirms PEFT-enhanced Whisper models as viable for equitable ASR in Tamil.

References

Kartik Addanki, John J. Godfrey, and Sanjeev Khudanpur. 2022. Acce: A benchmark for evaluating asr robustness to accent variations. In *Proceedings of Interspeech*.

Samuel Albanie, Gül Varol, Liliane Momeni, and 1 others. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. *European Conference on Computer Vision (ECCV)*.

Saleema Amershi, Daniel Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, and Penny Collisson. 2019. Guidelines for human-ai interaction. In *CHI Conference on Human Factors in Computing Systems*.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Rajalakshmi R, Suhasini S, and Swetha Valli. 2025. Overview of the fourth shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, Naples. Association for Computational Linguistics.

Ryan Baker, Yi Zhang, Zach Traylor, Mohit Doss, and Kristen Shinohara. 2020. Evaluating the effectiveness of speech recognition for individuals with speech impairments. In *ASSETS*.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–766, Toronto, Canada. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT*, pages 610–623.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Debasmita Bose, Naveena Karusala, and Denzil Ferreira Chattopadhyay. 2019. Voice as agency: Gender identity in voice user interfaces. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*.

Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. 2018. A comparison of audio signal preprocessing methods for deep neural networks on music tagging. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, pages 1870–1874. EURASIP.

Heidi Christensen. 2013. Automatic speech recognition for disordered speech. In *Handbook of Speech Communication*, pages 549–566. Walter de Gruyter.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of Interspeech 2021*, pages 2426–2430.

John R Deller, John H L Hansen, and John G Proakis. 1993. *Discrete-time Processing of Speech Signals*. Macmillan Publishing Company.

Emmanuel Deruty, Luc Leroy, Yann Macé, and David Meredith. 2025. Methods for pitch analysis in contemporary popular music: Highlighting pitch uncertainty in primaal's commercial works. *arXiv preprint arXiv:2502.08131*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Tiago H. Falk and William Y. Chan. 2007. Performance measures for voice-controlled interfaces with limited training data. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1169–1179.

Lorenzo Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. A survey on computational limitations and challenges in end-to-end speech recognition. *Computer Speech & Language*, 67:101178.

Mark Gales, Kate Knill, and Phil Woodland. 2019. Speech technology for health care: Opportunities and challenges. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6825–6829.

Ben Gold and Nelson Morgan. 2000. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. In *Proc. Interspeech*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, and 1 others. 2021. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICASSP*, pages 7383–7387. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.

Hye jin Shim, Md Sahidullah, Jee weon Jung, Shinji Watanabe, and Tomi Kinnunen. 2024. Beyond silence: Bias analysis through loss and asymmetric approach in audio anti-spoofing. *arXiv preprint arXiv:2406.17246*.

Daniel Jurafsky and James H Martin. 2000. *Speech and Language Processing*. Prentice Hall.

Hamza Kheddara, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *arXiv preprint arXiv:2403.01255*.

Zulfidin Khodzhaev. 2024. A practical guide to spectrogram analysis for audio signal processing. *arXiv preprint arXiv:2403.09321*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Shu Liu and Wayne G. Lutters. 2021. Speech interfaces for older adults: Supporting aging in place. *ACM Transactions on Accessible Computing (TACCESS)*, 14(3):1–26.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-tuning asr models for very low-resource languages: A study on mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Anas Malik, Zubayer Hossain, and Firoj Alam. 2021. Automatic speech recognition: A review. In *2021 2nd International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pages 1–6. IEEE.

Ziqiang Meng, Shuai Li, Dong Yu, and 1 others. 2021. A survey on speech evaluation metrics. *APSIPA Transactions on Signal and Information Processing*, 10:1–14.

Alan Morris, Victor Maier, and Phil Green. 2004. Spoken language understanding: Systems for extracting semantic information from speech. *IEEE Signal Processing Magazine*, 21(5):67–76.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *Interspeech*, pages 2616–2620.

Shrikanth Narayanan and Panayiotis Georgiou. Real-time emotion detection from speech using audio and voice quality features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Shweta Pandey and Akhilesh Mishra. 2019. Transgender inclusion and voice technology: A social justice perspective. *Technology in Society*, 59:101–120.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. 2021. End-to-end speech recognition: A survey. *arXiv preprint arXiv:2108.10520*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, and 1 others. 2020. Mls: A large-scale multilingual dataset for speech research. *Interspeech*.

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yan Chang, and Tara N. Sainath. 2019. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Lawrence R Rabiner and Ronald W Schafer. 1978. Digital processing of speech signals. *Prentice-Hall Signal Processing Series*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Alec Radford, Jong Wook Kim, Tian Xu, Greg Brockman, and Christine McGuffie. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 19123–19139. PMLR.

C Ramesh and Ramesh Gupta. 2021. Challenges in speech recognition for low-resource and regional indian languages. In *IEEE International Conference on Communication and Signal Processing*, pages 145–150. IEEE.

Javier Ramírez, Juan M Górriz, and Juan C Segura. 2007. Voice activity detection. fundamentals and speech recognition system robustness. *Robust speech recognition and understanding*, 1:1–22.

Ramya Srinivasan, Shalini Aravindhan, and Ravi Mahalingam. 2023. Inclusive speech recognition: Annotating transgender and elderly tamil voices. In *Proceedings of LT-EDI*.

Tuomas Tolonen and Matti Karjalainen. 2000. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8(6):708–716.

Werner Verhelst and Maarten Roelands. 2000. Perceptual audio coding based on a signal-adaptive psychoacoustic model. *IEEE Transactions on Speech and Audio Processing*, 8(3):330–338.

Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Guorong Xuan, Wei Zhang, and Peiqi Chai. 2001. Em algorithms of gaussian mixture model and hidden markov model. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2001)*, pages 733–736. IEEE.

Chenglin Yu, Ming Yin, Shuai Wang, Zhenhua Chen, and Xiaodong Wang. 2021. M2met: A multi-modal multi-genre dataset for speaker profiling and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3730–3743.

Dong Yu and Li Deng. 2017. Automatic speech recognition. In *Automated Speech Recognition*, pages 1–18. Springer.

Yu Zhang, Yashesh Wu, William Chan, Navdeep Jaitly, and Quoc V. Le. 2021. Benchmarking robust speech recognition: Background noise, babble, and channel variation. In *Proc. Interspeech*, pages 3076–3080.

SKV trio@LT-EDI-2025: Hybrid TF-IDF and BERT Embeddings for Multilingual Homophobia and Transphobia Detection in Social Media Comments

Konkimalla Laxmi Vignesh [1], Mahankali Sri Ram Krishna[1], Dondluru Keerthana[1], Premjith B[1],
[1]Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidyapeetham, India
Correspondence: b_premjith@cb.amrita.edu

Abstract

This paper presents a description of the paper submitted to the Shared Task on Homophobia and Transphobia Detection in Social Media Comments, LT-EDI at LDK 2025. We propose a hybrid approach to detect homophobic and transphobic content in low-resource languages using Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT) for contextual embeddings. The TF-IDF helps capture the token's importance, whereas BERT generates contextualized embeddings. This hybridization subsequently generates an embedding that contains statistical surface-level patterns and deep semantic understanding. The system uses principal component analysis (PCA) and a random forest classifier. The application of PCA converts a sparse, very high-dimensional embedding into a dense representation by keeping only the most relevant features. The model achieved robust performance across eight Indian languages, with the highest accuracy in Hindi. However, lower performance in Marathi highlights challenges in low-resource settings. Combining TF-IDF and BERT embeddings leads to better classification results, showing the benefits of integrating simple and complex language models. Limitations include potential feature redundancy and poor performance in languages with complex word forms, indicating a need for future adjustments to support multiple languages and address imbalances.

1 Introduction

Homophobia and transphobia are two concepts that foster negative opinions about homosexual and transgender individuals (Chakravarthi et al., 2022). These concepts are endemic in online communities and are most commonly expressed through context-dependent, subtle language that excludes LGBTQ+ populations. As digital communication grows, automated systems to detect and mitigate negative content using discriminatory language become increasingly important. However, detecting such content in low-resource languages is a significant challenge posed by linguistic subtleties and the lack of available annotated data.

This paper addresses the submission of the team SKV trio to the shared task on Homophobia and Transphobia Detection in Social Media Comments LT-EDI at LDK 2025 (Kumaresan et al., 2025). The proposed system was designed as a classification model with hybrid features. The proposed model combines features obtained from Term Frequency-Inverse Document Frequency (TF-IDF) and a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model. Integrating these categories of features, the system achieves both surface patterns and deeper contextual meaning required for efficient classification. The classifier was trained using a random forest classifier with the combined feature representations. We employed Principal Component Analysis (PCA) to facilitate the reduction of dimensionality within the TF-IDF embedding to match the TF-IDF and BERT embedding dimensionality. Our experiments demonstrate that the fusion of TF-IDF and BERT embeddings significantly improves classification accuracy compared to using either feature extraction technique alone. The implementation is available at https://github.com/K-LAXMIVIGNESH/SKVtrio-/tree/main

2 Related Work

The study (Chakravarthi et al., 2022) to detect homophobia and transphobia makes the hypothesis that multilingual language models can detect hate speech on social media more effectively if data augmentation through pseudolabeling is used. To test the hypothesis, the study looked at a number of multilingual language models. With 5,193 Malayalam and 3,203 Hindi comments, Kumare-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 26–30

san P et al. (Kumaresan et al., 2023) offer a new, high-quality dataset for identifying homophobic and transphobic content in both languages. The dataset also contains tests on the Malayalam, Hindi, English, Tamil, and Tamil-English datasets using transformer-based deep learning models and conventional machine learning. To automatically identify homophobic and transphobic content, (Chakravarthi, 2024) proposed a new dataset that was annotated by experts and trained machine learning models using it. 15,141 annotated comments in Tamil, English, and both languages are included in the dataset. The average macro F1 score for the top Tamil, English, and Tamil-English systems was 0.570, 0.870, and 0.610, respectively. These scores were obtained by training a random forest with fastText for Tamil data and logistic regression with BERT features for English and Tamil-English data. (Kumaresan et al., 2024) proposed a new expert-labeled dataset for Telugu, Kannada, and Gujarati, along with an expert-labeled dataset. The dataset includes extensive annotation rules, gathering about 10,000 comments in all three languages. A baseline model with pre-trained transformers was trained using the proposed dataset. The findings of a shared task to detect homophobic and transphobic language in social media posts are presented in (Chakravarthi et al., 2023) and (Chakravarthi et al., 2024). Task B required a fine-grained seven-class classification, while Task A classified comments into homophobic, transphobic, or neither categories. English, Spanish, Tamil, Hindi, and Malayalam were the five languages in which the task contained data. With the highest F1-scores of 0.997 for Spanish in Task A and 0.884 for Tamil in Task B, the top systems showed excellent performance. All the models reported in the literature are using either conventional techniques or BERT-based models for generating features. BERT-based models may not generate relevant features for rare words, which are important. These words can be efficiently captured by the TF-IDF algorithm.

3 Dataset description

The dataset is divided into training, development, and testing sets for each language. Table 1 shows the number of samples for each split.

There is noticeable variation in dataset size across languages. Kannada and Telugu, for instance, have significantly more samples than Hindi or Tamil, introducing language imbalance that can impact model generalization.

Table 1: Data distribution across languages

Language	Train	Dev	Test
English	3164	792	990
Gujarati	8119	1740	1740
Hindi	2560	320	321
Kannada	10063	2157	2156
Malayalam	3114	1213	866
Marathi	3500	750	750
Tamil	2662	666	833
Telugu	9050	1940	1939

4 Methodology

This section outlines the complete methodology of the system designed for this task. The methodology is divided into the following stages: data preprocessing, feature extraction, embedding fusion, model training, and prediction. Figure 1 illustrates the machine learning pipeline we used for this shared task. Three different data sources support the model: training, development, and test data.

In the preprocessing step, we check for blank rows and columns and remove such rows and columns from the dataset. We also searched for and removed any rows and columns that lacked text or labels. Finally, we converted all the labels into integers for processing.

We passed the preprocessed data through two different paths to extract two different types of features. We used the TF-IDF and BERT algorithms for feature extraction. TF-IDF captures the importance of words and tokens based on their frequency in the corpus. The TF-IDF feature extraction module was followed by an PCA module for dimensionality reduction. This algorithm uses 42 principal components for sampling features from a relatively lower dimensional space. BERT extracts contextual information from the input sequence and transforms the input text into a vector representation. Here, we used MuRIL (Sreelakshmi et al., 2024) for generating the embeddings. MuRIL is trained over translated and transliterated data in addition to the text in the original script. This motivated us to use MuRIL embeddings for this task.

We combine these two dense and lower-dimensional representations in a feature fusion step. We concatenated the contextual embeddings and dimensionality-reduced TF-IDF embeddings in this

step. We performed a 5-fold cross-validation to ensure the model's robustness. Later, we trained the model using the training data and evaluated its performance using the development data. Finally, we used the model to predict the labels of the test data.

5 Experimental Results

We evaluated our hybrid model on eight Indian languages using average F1-score and accuracy as the primary metrics. The results indicate that the model performs consistently well across most languages, especially in high-resource conditions.

The best performance was observed for Hindi, followed by English and Malayalam. These results suggest that the model effectively captures linguistic features and contextual nuances in these languages.

Table 2: Average accuracy and F1-score across languages computed for validation data

Language	Accuracy	F1-score
Hindi	0.9465	0.9205
English	0.9425	0.9157
Malayalam	0.9249	0.9187
Gujarati	0.9018	0.9022
Telugu	0.8977	0.8981
Tamil	0.8674	0.8442
Kannada	0.8587	0.8593
Marathi	0.7394	0.6340

The performance of the proposed system on the test data is shown in Figure 3.

Table 3: Average accuracy and F1-score across languages computed for test data

Language	F1-score
Hindi	0.3260
English	0.3350
Malayalam	0.3960
Gujarati	0.8570
Telugu	0.8660
Tamil	0.3710
Kannada	0.8120
Marathi	0.2940

From the results, it is evident that the model overfits on all languages except Gujarati, Kannada, and Telugu, showing strong generalization capabilities, even in moderately resourced or code-mixed environments. Marathi had the lowest F1-score among all languages, indicating possible challenges such as limited data, quality issues, or class imbalance. Kannada showed slightly lower results compared to others, but still maintained a reasonable performance.

Overall, the combination of TF-IDF and BERT embeddings has proven effective for multilingual hate speech classification. These results highlight the need for further work on handling dataset imbalance and enhancing model performance for low-resource languages.

The hyperparameters used for building the model are listed in Table 4.

Hyperparameters	Value
TF-IDF max_features	5000
dim(reduced TF-IDF)	512
RBF Sampler gamma	1.0
BERT truncation	True
BERT padding	True
BERT max_length	512
Random forest n_estimators	100
Random forest criterion	Gini
Random forest max_features	sqrt

Table 4: Hyperparameters used for building the model

6 Conclusion

In this work, we present a hybrid method that combines TF-IDF feature extraction along with BERT-based contextual embeddings for multilingual transphobia and homophobia detection in Indian languages. Our work demonstrates the effectiveness of combining both shallow lexical and deep contextual representations to achieve efficient classification. The experimental outcomes indicate that the hybrid model is both strong at capturing surface patterns and deeper semantic meanings required to detect subtle hate speech.

7 Limitations

The fusion of TF-IDF and BERT features can lead to feature redundancy or overfitting, as TF-IDF captures frequency-based semantics while BERT captures context, potentially introducing conflicting signals. We used the general BERT model for feature extraction, which may cause underperformance on low-resource or morphologically rich languages, especially if no multilingual fine-tuning is done.

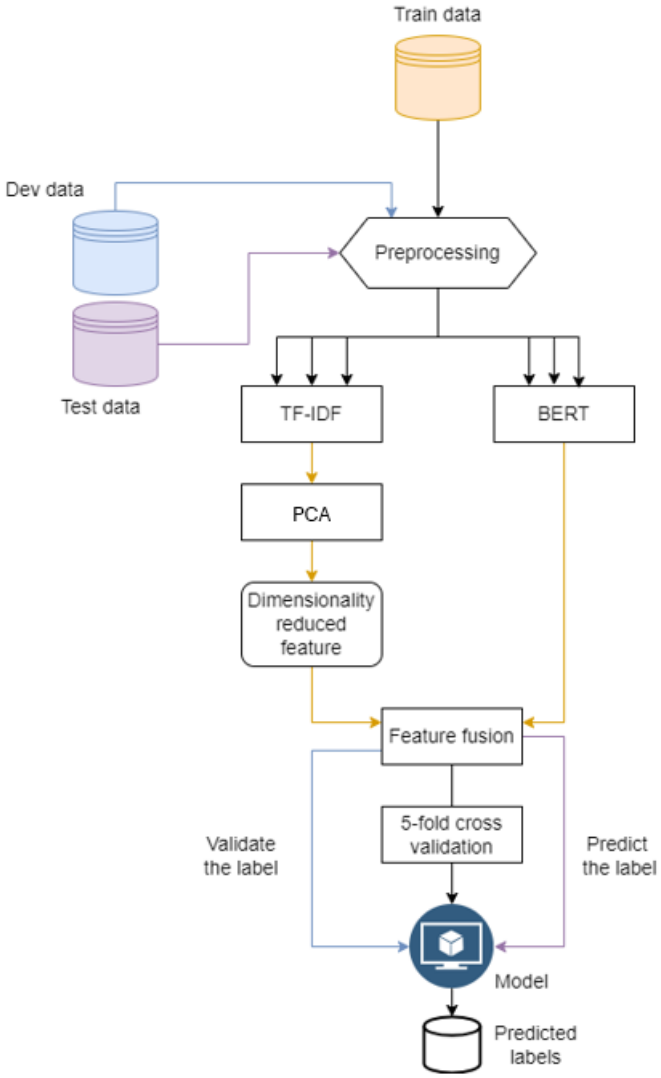

Figure 1: The block diagram explaining the workflow

References

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, 18(1):49–68.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, and 1 others. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Nitesh Jindal, and 1 others. 2023. Overview of second shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 38–46.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Paul Buitelaar, Malliga Subramanian, and Kishore Kumar Ponnusamy. 2025. Overview of fourth shared task on homophobia and transphobia span detection in social media comments. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and transphobia

detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, 5:100041.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. Dataset for identification of homophobia and transphobia for telugu, kannada, and gujarati. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

Dll5143A@LT-EDI 2025: Bias-Aware Detection of Racial Hoaxes in Code-Mixed Social Media Data (BaCoHoax)

Ashok Yadav and Vrijendra Singh
Indian Institute of Information Technology Allahabad, Prayagraj, 211015, India
{rsi2021002,vrij}@iiita.ac.in

Abstract

The proliferation of racial hoaxes that associate individuals or groups with fabricated crimes or incidents presents unique challenges in multilingual social media contexts. This paper introduces BaCoHoax, a novel framework for detecting race-based misinformation in code-mixed content. We address this problem by participating in the "Shared Task Detecting Racial Hoaxes in Code-Mixed Hindi-English Social Media Data: LT-EDI@LDK 2025." BaCoHoax is a bias-aware detection system built on a DeBERTa-based architecture, enhanced with disentangled attention mechanisms, a dynamic bias discovery module that adapts to emerging narrative patterns, and an adaptive contrastive learning objective. We evaluated BaCoHoax on the HoaxMixPlus corpus, a collection of 5,105 YouTube comments annotated for racial hoaxes, achieved a macro F1 score of 0.67, and secured 7th place among participating teams in the shared task. Our findings contribute to the growing field of multilingual misinformation detection and highlight the importance of culturally informed approaches to identifying harmful content in linguistically diverse online spaces.

1 Introduction

The proliferation of social media platforms has transformed how information spreads across diverse linguistic communities, creating both opportunities and challenges for marginalized populations worldwide ((Gowen et al., 2012); (Yates et al., 2017)). During recent global events, including the COVID-19 pandemic, online platforms have become critical spaces for emotional expression and support seeking (Wang and Jurgens, 2018). For vulnerable communities such as women in STEM, LGBTIQ individuals, racial minorities, and people with disabilities. These digital spaces significantly influence self-perception and societal integration (Chung, 2014); (Altszyler et al., 2018); (Tortoreto et al., 2019). While numerous studies have focused on detecting negative content through hate speech recognition (Schmidt and Wiegand, 2017), offensive language identification (Yadav et al., 2024), and implicit hate detection (Yadav and Singh, 2024), these approaches often fail to address underlying biases and may inadvertently discriminate against minority groups. The spread of misinformation targeting specific communities presents a particularly concerning challenge, especially in multilingual contexts where traditional detection systems struggle to operate effectively. This is especially problematic in code-mixed environments where individuals switch between multiple languages within single utterances (Chakravarthi, 2020), creating linguistic patterns that evade conventional detection methods. While research has expanded to include languages beyond English, including Arabic, German, Hindi, and Italian, these studies primarily examine monolingual corpora, overlooking the complexity of code-switched communication prevalent in diverse linguistic regions like South Asia.

The key contributions of our work include: (1) a DeBERTa-based architecture enhanced with disentangled attention mechanisms specifically optimized for code-mixed content; (2) a Dynamic Bias Discovery component that continuously identifies potentially biased terms throughout the training process, adapting to emerging narrative patterns; and (3) an adaptive contrastive learning approach that enhances the model's ability to distinguish subtle patterns in misinformation. Through comprehensive evaluation and error analysis, we demonstrate both the effectiveness of our approach. Code: [1]

[1]Code: https://github.com/ashokiiita/LDK_2025

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 31–38

2 Task and Dataset Description

The LT-EDI@LDK 2025 Shared Task focused on the detection of racial hoaxes in code-mixed Hindi-English content, addressing the critical challenge of misinformation that targets specific social or ethnic communities (Chakravarthi et al., 2025). The primary objective of the shared task was to develop computational systems capable of automatically identifying such content in low-resource, multilingual contexts. Participating teams were required to build binary classification systems that could distinguish between regular content (non-hoax) and racial based misinformation (hoax) in code-mixed text, where speakers switch between Hindi and English within the same utterance. The task evaluation employed macro-averaged F1 score as the primary metric, which equally weights performance on both the majority (non-hoax) and minority (hoax) classes, thereby encouraging systems to effectively identify the less frequent but more harmful racial hoax content.

The task organizers provided the HoaxMixPlus, a corpus comprising 5,105 YouTube comment annotated for racial hoaxes, as shown in Table 1. The HoaxMixPlus dataset is distributed across train (3,060), validation (1,021), and test (1,021) (Chakravarthi, 2020).

Table 1: Statistics of the dataset used in the LT-EDI@LDK 2025 Shared Task

Dataset	Total Samples	Non-Hoax (0)	Hoax (1)
Train	3060	2319	741
Validation	1021	774	247
Test	1021	774	247

3 Proposed Methodology

In this section, we present BaCoHoax, a novel framework for detecting racial hoaxes in code-mixed content. Our approach extends the capabilities of pretrained language models by integrating bias-aware components specifically designed to identify and leverage linguistic patterns associated with hoaxes. The framework consists of four main components: (1) a DeBERTa-based encoder, (2) a dynamic bias discovery mechanism, (3) disentangled bias-aware attention, and (4) a contrastive learning objective.

We formulate the task as a supervised classification problem. Given a text input $X = \{x_1, x_2, ..., x_n\}$ consisting of n tokens, our goal is to predict whether it contains a racial hoax, represented as a binary label $y \in \{0, 1\}$, where 1 indicates the presence of Hoax.

3.1 Base Architecture: DeBERTa-based encoder

Our model employs DeBERTa-v3 (He et al., 2021) as the backbone encoder due to its enhanced disentangled attention mechanism, which effectively separates content and positional information. This property is particularly valuable for our task, as it allows the model to better capture subtle contextual relationships in code-mixed content where sensitive terms may appear in varying positions and contexts. The encoder maps each input token to a contextualized representation, producing a sequence of hidden states $H = \{h_1, h_2, ..., h_n\}$. To capture richer semantic information, we employ a feature fusion approach that combines information from multiple layers using equation 1:

$$F = \text{LayerNorm}(W_f[H^L \oplus H^{L-1} \oplus H^{L-2}] + b_f) \quad (1)$$

where H^L, H^{L-1}, and H^{L-2} are the hidden states from the last three layers of DeBERTa, $\oplus$ denotes concatenation, and W_f and b_f are learnable parameters. This multi-layer fusion allows the model to leverage both high-level abstract features and lower-level syntactic information.

3.2 Dynamic Bias Discovery

In our approach we have used the DBD component, which continuously identifies potentially biased terms throughout the training process. Unlike static approaches that rely on predefined lexicons, DBD learns to recognize bias patterns directly from the data.

i. **Identity Term Identification:** We initialize the system with a seed list of common identity terms (e.g., religious, caste, and regional identifiers) in Hindi-English code-mixed text. For each input text, DBD identifies tokens that match predefined identity patterns.

ii. **Class Distribution Tracking:** For each identified term, DBD maintains a distribution of occurrences across the target classes (misinformation vs. factual), enabling the calculation of bias scores using equation 2.

$$\text{BiasScore}(t) = \frac{\text{count}(t, y = 1)}{\text{count}(t, y = 0) + \text{count}(t, y = 1)} \quad (2)$$

iii. **Contextual Embedding:** For each bias term, DBD stores and updates representative embeddings by averaging the contextualized representations of the term across all its occurrences.

iv. **Threshold-based Selection:** Terms with strong class associations (bias scores significantly above or below 0.5) and sufficient occurrence count are added to the bias term set.

This component allows our model to adaptively discover new bias terms without manual annotation, making it robust to evolving language patterns and novel bias expressions.

3.3 Disentangled Bias Detection and Attention

To effectively leverage the discovered bias terms, we implement two specialized components:

3.3.1 Disentangled Bias Detector

The DBD identifies potential bias signals in the input text by comparing token embeddings with the embeddings of known bias terms. For each token in the input sequence, DBD computes a bias probability by identifying tokens that match known bias terms, then extracting a context window around each identified token. It subsequently applies disentangled attention to evaluate the contextual usage of each term, and finally computes a bias probability for the token along with its surrounding context. This results in a bias mask $M_{bias} \in \mathbb{R}^n$ that highlights tokens likely to contribute to biased or misleading content.

3.3.2 Disentangled Bias Attention

The Disentangled Bias Attention (DBA) module extends DeBERTa's disentangled attention mechanism by incorporating bias awareness:

$$A_{i,j} = \frac{(W_Q h_i)^T (W_K h_j)}{\sqrt{d}} + \alpha \cdot M_{bias}[j] \quad (3)$$

where W_Q and W_K are query and key projection matrices, d is the dimension of the projection, and α is a learnable parameter that controls the influence of the bias mask. This attention mechanism amplifies focus on tokens identified as potentially biased, allowing the model to better analyze their contextual usage.

The bias-aware context representation is computed using equation 4:

$$C_{bias} = \text{softmax}(A) \cdot (W_V H) \quad (4)$$

This representation is then combined with the standard sentence representation to form a comprehensive feature vector.

3.4 Classification and Learning Objectives

Our model employs a multi-objective learning approach that combines classification and contrastive learning:

The classification head takes the concatenated representation of the [CLS] token embedding and the bias-guided vector:

$$Z = [h_{CLS} \oplus v_{bias}] \quad (5)$$

This is passed through a multi-layer classifier to produce class logits:

$$\hat{y} = \text{softmax}(W_c \text{MLP}(Z) + b_c) \quad (6)$$

where MLP is a multi-layer perceptron with layer normalization and GELU activation functions.

To enhance the model's ability to distinguish subtle patterns in hoax content, we implemented an adaptive contrastive learning objective. For each input sample, we create an augmented version using DeBERTa-specific augmentation techniques such as Entity emphasis (duplicating entity mentions), Word deletion (removing non-essential words) and Synonym replacement (substituting words with similar meanings).

These enhancements preserve semantic content while creating variations that help the model learn robust representations. The contrastive loss is computed using an adaptive temperature scaling approach using equation 7.

$$\mathcal{L}_{cont} = -\frac{1}{|P|} \sum_{(i,j) \in P} \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(z_i \cdot z_k / \tau)} \quad (7)$$

where P is the set of positive pairs, z_i and z_j are normalized projection embeddings, and τ is an adaptive temperature parameter adjusted based on batch similarity distribution. The final loss combines classification and contrastive objectives using equation 8.

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{cont} \quad (8)$$

where λ is dynamically adjusted during training based on validation performance, allowing the model to find an optimal balance between the two objectives.

4 Experimental Settings and Result Analysis

We implement our model using PyTorch and the Transformers library. For the DeBERTa encoder, we use the `microsoft/deberta-v3-base` variant with 12 layers and a hidden size of 768. We set the maximum sequence length to 128 tokens and use a batch size of 8 for training. The AdamW optimizer is configured with a base learning rate of 2×10^{-6} and a maximum learning rate of 1×10^{-5}, with a weight decay of 0.01. We train the model for 10 epochs with early stopping based on the F1 validation score. For the bias discovery component, we use a bias threshold of 0.65 and a minimum occurrence count of 5. The contrastive learning weight is initialized at 0.15 and dynamically adjusted during training within the range [0.05, 0.3]. All experiments were performed using a NVIDIA A30 GPU.

The performance of our BaCoHoax framework was evaluated in the context of the shared task, where systems were primarily ranked based on their macro-averaged F1-scores. Table 2 summarizes the official results and standings of all participating teams.[2]

Table 2: Official Ranking of Teams in the HoaxMixPlus Shared Task

Team Name	Macro F1 Score	Rank
CUET's_White	0.75	1
Hope_for_best	0.72	2
KCRL	0.71	3
HoaxTerminator	0.70	4
Hinterwelt	0.69	5
Belo Abhigyan	0.68	6
KEC-Elite-Analytics	0.68	6
Dll5143A	**0.67**	**7**

Our BaCoHoax model achieved a macro F1 score of 0.67, securing 7th position among all participating teams. This performance demonstrates the effectiveness of our approach in addressing the challenging task of detecting racial hoaxes in code-mixed content. The top-performing system (CUET's_White) achieved a macro F1 score of 0.75, The relatively tight clustering of scores among the top 8 teams (ranging from 0.75 to 0.63) suggests that detecting racial hoaxes in code-mixed Hindi-English content remains challenging, with multiple approaches achieving similar performance levels. The detailed results are discussed in the Appendix A.1

[2]Official leaderboard: https://codalab.lisn.upsaclay.fr/competitions/21885#learn_the_details-evaluation

5 Conclusion and Future Work

In this paper, we introduced BaCoHoax, a novel framework for detecting race based Hoax content in code-mixed content that leverages disentangled attention mechanisms and bias-aware representations. Our approach incorporates cultural and linguistic nuances through a dynamic bias discovery mechanism and contextual understanding of bias terms in multilingual settings. We achieved a competitive macro F1 score of 0.67 and securing 7th place among participating teams in the shared task. The pre-defined bias term approach creates problematic overgeneralization and identity-bias conflation, where the model cannot effectively distinguish between merely mentioning an identity group and expressing bias toward that group.The imapct of pre-deifned bias terms are discussed in detail A.4. These findings suggest the need for context-aware and culturally grounded bias detection methods, along with improved discourse analysis and refined lexicon strategies. Future models should also consider multi-stage architectures to better handle the complexity of code-mixed and nuanced biased content.

6 Error Analysis

We conducted detailed error analysis which provides insights of specific areas where our model faced difficulties, particularly in detecting subtle forms of racial bias expressed through cultural references, rhetorical devices, and implicit language challenges. The detailed error analysis is dicussed in Appendix A.2. results validate the effectiveness of our BaCoHoax framework while highlighting opportunities for further refinement in future iterations.

References

Edgar Altszyler, Ariel J Berenstein, David N Milne, Rafael A Calvo, and Diego Fernandez Slezak. 2018. Using contextual information for automatic triage of posts in a peer-support forum. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 57–68.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Jae Eun Chung. 2014. Social networking in online support groups for health: how online social networking benefits patients. *Journal of health communication*, 19(6):639–659.

Kris Gowen, Matthew Deschaine, Darcy Gruttadara, and Dana Markey. 2012. Young adults with mental health conditions and social networking websites: seeking tools to build community. *Psychiatric rehabilitation journal*, 35(3):245.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Giuliano Tortoreto, Evgeny A Stepanov, Alessandra Cervone, Mateusz Dubiel, and Giuseppe Riccardi. 2019. Affective behaviour analysis of on-line user interactions: Are on-line support groups more therapeutic than twitter? *arXiv preprint arXiv:1911.01371*.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45.

Ashok Yadav, Farrukh Aslam Khan, and Vrijendra Singh. 2024. A multi-architecture approach for offensive language identification combining classical natural language processing and bert-variant models. *Applied Sciences*, 14(23):11206.

Ashok Yadav and Vrijendra Singh. 2024. Hatefusion: Harnessing attention-based techniques for enhanced filtering and detection of implicit hate speech. *IEEE Transactions on Computational Social Systems*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

A Appendix

A.1 Appendix A: Result Analysis

Table 3 presents the detailed performance metrics of the BaCoHoax model on the training set. The model achieves an overall accuracy of 86%, indicating that it correctly classifies approximately four out of five instances in the training data. However, this aggregate metric masks significant disparities in the model's ability to detect different classes of content.

For non-racial content (class 0), the model demonstrates strong performance with a precision of 0.88 and recall of 0.93, resulting in an F1-score of 0.91. This indicates that when the model predicts content as non-racial, it is correct 88% of the time, and it successfully identifies 93% of all non-racial content in the dataset. These metrics suggest that the model has developed a robust understanding of non-racial content patterns. In contrast, the model's performance on racial content (class 1) is considerably less stronger. With a precision of 0.75 and recall of only 0.61, the resulting F1-score is a modest 0.67. The macro-average metrics (precision: 0.81, recall: 0.77, F1-score: 0.79) provide a class-balanced view of performance and highlight the significant room for improvement in identifying racial content. The weighted averages (precision: 0.76, recall: 0.79, F1-score: 0.77) appear somewhat higher due to the dominance of non-racial samples in the dataset and the model's stronger performance on this majority class. These training set metrics highlight a core challenge in the bias detection task. While the model performs well in identifying non-racial content, it struggles to detect racial bias, particularly when it is expressed in subtle or implicit forms.

Table 3: Training Set Performance Metrics of the BaCoHoax Model

Class	Precision	Recall	F1-Score
Non-Racial (0)	0.88	0.93	0.91
Racial (1)	0.75	0.61	0.67
Accuracy		0.86	
Macro Average	0.81	0.77	0.79
Weighted Average	0.85	0.86	0.85

Table 4 presents the performance metrics of our BaCoHoax model on the validation set. The model overall accuracy of 75%, indicating good performance on unseen data. For non-racial content (class 0), the model demonstrates robust performance

with a precision of 0.83 and an exceptionally high recall of 0.85, resulting in an F1-score of 0.84. This indicates that the model effectively recognizes 85% of all non-racial content in the validation set, and when it classifies content as non-racial, it is correct 83% of the time. These results confirm that the model has developed a strong capacity to identify non-racial content patterns that generalize well to unseen data.

However, the model's performance on racial content (class 1) shows significant weaknesses. With a precision of 0.49 and a very low recall of 0.45, the resulting F1-score is only 0.47. These metrics reveal a critical limitation in the model's ability to detect racially-charged content. The macro-average F1-score of 0.65 provides a class-balanced performance metric that emphasizes the substantial performance gap between classes. The weighted average F1-score of 0.75 appears higher primarily due to the class imbalance in the dataset and the model's stronger performance on the majority class.

The validation results reveal a persistent challenge in racial content detection, while the model generalizes well for non-racial content. The lower recall on racial content compared to the training set suggests that the model may be encountering forms of racial bias in the validation set that differ from those in the training data, highlighting the difficulty of capturing the diverse and evolving expressions of racial bias in code-mixed text.

Table 4: Validation Set Performance Metrics of the BaCoHoax Model

Class	Precision	Recall	F1-Score
Non-racial (0)	0.83	0.85	0.84
Racial (1)	0.49	0.45	0.47
Accuracy		0.75	
Macro Average	0.66	0.65	0.65
Weighted Average	0.75	0.75	0.75

Table 5 presents the final performance metrics of our BaCoHoax model on the held-out test set. The model achieves an overall accuracy of 76%. For non-racial content (class 0), the model exhibits strong performance with a precision of 0.85 and recall of 0.84, resulting an F1-score of 0.84. These metrics show that the model effectively identifies 84% of all non-racial content in the test set, and when it classifies content as non-racial, it is correct 85% of the time.

However, the model's performance on racial content (class 1) continues to show substantial limitations. With a precision of 0.51 and a recall of 0.53, the resulting F1-score is 0.52. These figures indicate that the model fails to identify 47% of racial content in the test data, and when it does predict content as racial, it is incorrect 49% of the time. The macro-average F1-score of 0.67 and weighted average F1-score of 0.77 are consistent with those observed in the training set, confirming the model's stable performance pattern across dataset. The significant gap between non-racial and racial content detection capabilities remains a persistent characteristic of the model. The test results confirm that while the BaCoHoax approach provides a solid foundation for racial bias detection, significant improvements are still needed to address the challenge of identifying subtle and implicit forms of racial bias.

Table 5: Test Set Performance Metrics of the BaCoHoax Model

Class	Precision	Recall	F1-Score
Non-racial (0)	0.85	0.84	0.84
Racial (1)	0.51	0.53	0.52
Accuracy		0.76	
Macro Average	0.68	0.68	0.67
Weighted Average	0.77	0.76	0.77

A.2 Appendix

B: Error Analysis

A.2.1 Analysis of False Positive Errors

We analyzed instances where our model incorrectly classified non-racial content as containing racial bias. Table 6 presents a selection of these false positive cases.

Our analysis of false positive cases reveals consistent patterns in the model's tendency to over-classify content as racially biased. These patterns underscore key limitations in the current approach to bias detection, particularly in the context of code-mixed text. One prominent issue is the model's frequent misinterpretation of general profanity and aggressive language as signals of racial bias. For instance, in example RH_10, the text criticizes the news media using terms like "chor gadar atankwadi" (thief, traitor, terrorist), yet it does not target any specific religious or ethnic group. Similarly, RH_30 includes profane insults such as "shale suar ki olad" but focuses on disrupting societal peace rather than targeting protected communities.

Table 6: Examples of False Positive Errors in the BaCoHoax Model

ID	Text Content	Predicted	True
RH_10	tum sare news walo chor gadar atankwadi ho jo sirf moh se zaher ugalna janti ho	1	0
RH_18	kiyo bat karte ho musalman ko pakistan jane ki esa hi hoga sarkar kadam nahi utaygi to majburan bolna padta hai kiyo army nahi bejte brma mei bacho ko kata ja rha hai ladies ki ijat loti ja rhi hai fast bola us par dyan diya nahi chutya kahi k ai hai bahsan dene k liye	1	0
RH_30	shale suar ki olad teri okat kya h ham desh me shanti chahte h barna sudhar jao salo	1	0
RH_35	congres party ke gadar desh drohi hai aise admi ko goli mardo madarchod sale ne bhrashtachar kiya hai...	1	0
RH_38	bihar in purani kaduai yado se age nikal chuka hai apne 10, 20 like coment ke like ye sab na kare to apne bihar ke like this hai	1	0
RH_51	modi ki hrkatu se yah sabit hua ki mohamad ali jinah thek tha aur abdul kalam azad ghalat hindustan ki musalmanon ka mujrim abdul kalam azad hai aur koi nai jis ne pakistan ki bajhe hindustan ka sathiya	1	0
RH_68	ak muslim hote hue me kh raha hu ke es truke ki jumla baji krne valo ko desh ka savidhan bade se bdi saja de...	1	0
RH_109	islam kaha se peda huva alah khush hoke peda kiya jatiwadi one jat ke nam logo ka shoshan atyachar...	1	0
RH_119	mujhe hindu musalman nahinsan aur insaniyat kahatam hote dikh rahi haialah har bande ko nek hidayat...	1	0
RH_129	to hinduo bijli pani fri ke liye kejriwal ko vote b tumne hi kiya tha ab jutey khao	1	0

These examples suggest that the model exhibits a problematic bias by linking profanity with race. A second pattern involves the misclassification of political criticism as racial bias. In RH_35, the model flags a harsh critique of the Congress party using the phrase "gadar desh drohi" (traitor, anti-national) and referencing corruption, even though there is no mention of race or religion. This indicates that the model may be over-associating political discourse with bias, possibly due to correlations in the training data where political references frequently co-occurred with biased content.

The model also struggles to distinguish between neutral or positive mentions of identity groups and actual bias. In RH_68, a self-identified Muslim user condemns divisive rhetoric and advocates for peace and harmony. Similarly, RH_119 explicitly states, "mujhe Hindu Musalman nahin, insaan aur insaniyat" (I don't see Hindu/Muslim, but humanity), yet the model still classifies the text as biased. These misclassifications suggest that the model relies heavily on the presence of identity terms without fully understanding their contextual meaning. Another common source of error involves historical or factual references to sensitive topics. In RH_51, the speaker discusses historical figures like Jinnah and Azad, while RH_109 explores the origins of Islam and caste-based oppression. Neither example promotes contemporary bias, yet both are flagged by the model.

We also observed that texts containing contentious arguments involving communities but lacking explicit bias which are often misclassified. RH_38 critiques the resurfacing of past divisions in Bihar, and RH_129 questions the political choices of Hindu voters without expressing religious animosity. These cases reflect a broader issue: the Disentangled Bias Detector tends to flag content containing identity-related keywords (e.g., "Hindu," "Muslim," "BJP," "Congress") regardless of context, especially when such terms frequently co-occurred with biased content during training. Similarly, references to historical conflicts are often treated as biased, even when discussed factually. These findings support our hypothesis that the use of predefined bias terms during model initialization may contribute to overgeneralization and the conflation of identity mention with bias. Limitations in contextual understanding are particularly evident in cases like RH_68 and RH_119, where the model detects religious terms but fails to recognize the underlying pro-harmony messages.

A.3 False Negative Analysis

Following our analysis of false positive cases, we conducted analysis of false negatives samples to further assess the limitations of the BaCoHoax model. Specifically, we analyzed instances where racially biased hoax content was incorrectly classified as non-biased. Table 7 presents representative examples of such misclassifications, illustrating the model's difficulty in recognizing subtle or contextually embedded racial bias.

Our analysis of these false negative cases reveals several consistent patterns in the model's inability to detect racial content. In example RH_07,

Table 7: Examples of False Negative Errors in the BaCoHoax Model

ID	Text Content	Predicted	True
RH_07	behan sambhal ke rahana in ka yeisa hai ki jab ap inki bat karoge ye thik hai par inki ashliyet batoge to tunhehi khatra paida karenge ye kisan nahi khalisthani hai...	0	1
RH_26	hamara desh ke khate mein police hamara hinduon ka tax do per hi atyachar hote hain yah kaun si bat ho gai ise banaoge hindu rashtra ke	0	1
RH_31	hasi ati h tere knowledge pe puri duniya me including india sab se bda inflict muslims me hi h	0	1
RH_40	ye khabar sunkar sch bta rha hu hath m talwar ho aur fir s unko katu jo in garibo ki aurat ko bhi nhi baksa aur inke jatiwad ko inki gand m dal deta fir dhekte h ki kon asli mard h	0	1
RH_48	teri soch kitni nech hai tere ma bap ki trha surnames kisi ke bap ki bapoti nhi hai jiska jo mn chahega wo wo likhega verma likhe ya harijan uski marzi aur shadi ke samay sb apni caste batate hai koi kuch nhi chupata lgta hai teri mumy ya pita ne aise hi shadi ki hogi bhag kr	0	1
RH_50	bhai ye musalman ki fake id hai delhi election mai in katwo ko pta chal jayega	0	1

the model fails to identify the implicit bias where protestors are characterized as "khalisthani" (a politically charged term) rather than legitimate farmers. Similarly, in RH_50, the model misses the derogatory reference to Muslims using the term "katwo," which is highly offensive in the given context. Many of these examples demonstrate the model's difficulty in recognizing coded language and cultural references that require deep understanding of the sociopolitical context.

For instance, RH_26 contains subtle implications about tax contributions and religious identity that require contextual understanding beyond simple term matching. The model also struggles with complex cases like RH_31 makes broad negative generalizations about Muslims that the model fails to detect, likely due to the absence of explicit slurs or recognized bias terms. Additionally, examples like RH_40 and RH_48 demonstrate the model's inability to identify violent rhetoric (references to weapons) and caste-based insults when they're embedded within colloquial expressions or aggressive language patterns common in social media discourse. These patterns highlight the need for more sophisticated approaches to bias detection that can capture subtle linguistic cues, cultural references, and implicit expressions of prejudice in code-mixed text. The evolving and evasive nature of biased language makes detection difficult.

A.4 Impact of Pre-defined Bias Terms

A critical factor in both false positive and false negative errors can be traced to the model's initialization with predefined bias terms. The BaCoHoax model was seeded with an initial set of bias terms including identity markers ("hindu", "muslim", "islam", "mandir", "masjid", "brahmin", "dalit", "sc", "st", "obc"), names associated with specific communities ("singh", "khan", "sharma", "ali", "kumar"), politically charged terms ("jihad", "bhakt", "sanghi", "liber"), and political party references ("congress", "bjp", "modi", "rahul"). This initial seeding creates several problematic effects that significantly impact model performance. The approach leads to overgeneralization, as the model is primed to flag any content containing these broad identity and political terms regardless of context. It also results in identity-bias conflation, where the model cannot effectively distinguish between merely mentioning an identity group and expressing bias toward that group, creating a fundamental "guilt by association" problem. Additionally, the inclusion of political terms as bias markers creates a problematic political-communal confusion, explaining many false positives in content that offers political criticism without communal bias. The model's bias detection mechanism demonstrates context-free processing, assuming these terms are inherently biased rather than recognizing their meaning is highly context-dependent. Furthermore, the Dynamic Bias Discovery component creates a self-reinforcing bias loop where neutral mentions may be incorrectly classified as biased, strengthening problematic associations over time. For false negatives, this approach fails to capture bias expressed through more subtle means without explicitly using the predefined terms, while for false positives, it overreacts to any content containing these terms regardless of context or intent.

Hope_for_best@LT-EDI 2025: Detecting Racial Hoaxes in Code-Mixed Hindi-English Social Media Data using a multi-phase fine-tuning strategy

Abhishek Singh Yadav[1] **Deepawali Sharma**[2] **Aakash Singh**[1] **Vivek Kumar Singh**[1]
[1]Department of Computer Science, University of Delhi, India
[2]School of Computer Science Engineering and Technology, Bennett University, Noida, India
meabhishek8965@gmail.com, deepawali21@bhu.ac.in
asingh@cs.du.ac.in, vivek@cs.du.ac.in

Abstract

In the age of digital communication, social media platforms have become a medium for the spread of misinformation, with racial hoaxes posing a particularly insidious threat. These hoaxes falsely associate individuals or communities with crimes or misconduct, perpetuating harmful stereotypes and inflaming societal tensions. This paper describes the team "Hope_for_best" submission that addresses the challenge of detecting racial hoaxes in code-mixed Hindi-English (Hinglish) social media content and secured the 2^{nd} rank in the shared task (Chakravarthi et al., 2025). To address this challenge, the study employs the HoaxMix-Plus dataset, developed by LT-EDI 2025, and adopts a multi-phase fine-tuning strategy. Initially, models are sensitized using the THAR dataset—targeted hate speech against religion (Sharma et al., 2024) —to adjust weights toward contextually relevant biases. Further fine-tuning was performed on the HoaxMix-Plus dataset. This work employed data balancing sampling strategies to mitigate class imbalance. Among the evaluated models, Hing-BERT achieved the highest macro F1-score of 73% demonstrating promising capabilities in detecting racially charged misinformation in code-mixed Hindi-English texts.

1 Introduction

In the digital era, social media platforms have revolutionized global communication by enabling individuals to disseminate information across vast and diverse audiences. However, this accessibility has also facilitated the rapid spread of misinformation, including racially charged hoaxes that falsely implicate individuals or communities in criminal or unethical behavior. Such hoaxes are not merely misinformative but are deliberately crafted to reinforce harmful stereotypes, incite hostility, and exacerbate societal divides (Singh et al., 2025).

Platforms like Twitter, Facebook, and YouTube empower users with the ability to express opinions publicly and anonymously. While this democratization of speech has positive implications, it also opens the door to misuse. These platforms, with their anonymous nature and rapid content spread, often intensify hate-fueled narratives. This makes it increasingly important to build automated systems that can identify and curb such content before it leads to real-world consequences (Shanmugavadivel et al., 2022).

In the domain of targeted hate speech detection, Natural Language Processing (NLP) has made significant progress with the advent of deep learning architectures (Sharma et al., 2025b). Early approaches employed Long Short-Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Chung et al., 2014), but the emergence of transformer-based models, particularly BERT (Vaswani et al., 2017), has significantly improved performance across a range of language understanding tasks.

One of the major challenges in this domain is dealing with code-mixed text, especially Hindi-English (Hinglish), which is commonly used in Indian social media discourse. This linguistic mixing complicates tokenization, syntactic parsing, and semantic understanding. While several studies have focused on hate speech detection in Indic languages (Mathew et al., 2021; Patwa et al., 2020), the problem of detecting racial hoaxes in Hinglish remains underexplored.

The present study contributes to this growing field by introducing a transformer-based approach for the classification of racial hoaxes in code-mixed Hindi-English social media content. Building on the HoaxMixPlus dataset introduced by LT-EDI 2025, this study employs a multi-phase fine-tuning strategy to adapt models for the detection of contextually biased misinformation. Initially, pretraining is conducted on the THAR dataset (Sharma et al., 2024), which targets hate speech against religious communities, followed by fine-tuning on

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 39–46

task-specific HoaxMixPlus data. Among the models evaluated, Hing-BERT achieved the best performance, demonstrating its effectiveness in capturing racially hostile content embedded in informal, code-mixed linguistic structures.

This paper is structured as follows: Section 2 provides a comprehensive review of the existing literature on misinformation and hate speech detection, specifically within the context of code-mixed languages. Section 3 outlines the datasets utilized in this study, namely the THAR and HoaxMixPlus datasets, along with a detailed description of their respective features. Section 4 describes the methodology employed, with an emphasis on the multi-phase fine-tuning approach, model architecture, and data preparation techniques. In Section 5, the experimental results are presented. Section 6 offers a detailed discussion and analysis of the model's performance. Finally, Section 7 concludes the paper and highlights potential avenues for future research.

2 Related Work

The detection of misinformation and racially motivated hoaxes in social media has attracted increasing attention in recent years, particularly in multilingual and code-mixed contexts. Prior studies have explored various linguistic and contextual challenges in identifying harmful narratives, such as hate speech, fake news, and racially biased misinformation.

There is a growing need for research addressing harmful and biased content in code-mixed and multilingual social media, supported by the creation of linguistically diverse datasets and model strategies. Studies such as HopeEDI (Chakravarthi, 2020) and the ensemble-based model for hope speech detection in English and Dravidian languages (Sharma et al., 2025a) highlight the effectiveness of such approaches in promoting inclusive and equitable language technologies.

Code-mixed language, especially Hindi-English (Hinglish), presents significant challenges for natural language understanding due to its informal structure and lack of standardized grammar. Recent efforts, such as Patwa et al. (2020), have addressed sentiment analysis and offensive language detection in code-mixed texts through SemEval-2020. Similarly, Barman et al. (2014) provided foundational insights into part-of-speech tagging in Bengali-Hindi-English code-mixed social media content, underscoring the complexity of such texts.

Although racial hoaxes remain an underexplored domain, prior research in hate speech detection provides valuable foundations. Datasets such as THAR (Sharma et al., 2024) target religious hate in multilingual Indian contexts, offering pretraining potential for models tackling similar sociolinguistic phenomena. Other notable works include Mathew et al. (2021), who introduced HateXplain—a benchmark dataset for hate speech detection with multi-perspective annotations including class labels, target communities, and human-provided rationales to improve model explainability and reduce bias, and Vidgen and Yasseri (2020), who developed a multi-class classifier to distinguish between non-Islamophobic, weakly Islamophobic, and strongly Islamophobic content, emphasizing the need for nuanced categorization over binary classification.

Transfer learning through sequential fine-tuning has shown considerable promise in improving task-specific performance for low-resource and domain-specific problems. The use of a multiphase fine-tuning pipeline, where models are initially exposed to related bias-aligned data (e.g., hate speech or religious hostility) and subsequently adapted to the target task (e.g., racial hoaxes), aligns with strategies explored in Gururangan et al. (2020), who demonstrated the efficacy of domain-adaptive pretraining. In the current work, models such as Hing-BERT leverage this approach by first calibrating on THAR before task-specific tuning on HoaxMixPlus.

Pretrained multilingual models like MuRIL (Kakwani et al., 2020) and domain-specific transformers such as Hing-BERT (Kumar et al., 2020) have been specifically optimized for Indian languages and their mixed variations. These models benefit from pretraining on diverse scripts and colloquial structures, making them suitable for nuanced detection tasks in Hinglish texts. Moreover, models like hing-roberta-mixed have demonstrated competitive performance in identifying hate speech in informal, noisy, and multilingual settings.

Given the skewed nature of real-world social media datasets, strategies like data sampling, loss re-weighting, and oversampling are commonly adopted to mitigate bias and improve minority class detection. Approaches documented in Rathpisey and Adji (2022) emphasize the importance of balancing in achieving fairer performance across all classes.

Despite the substantial progress in detecting hate

[1]https://www.aclweb.org/anthology/2020.peoples-1.5/

speech, fake news, and offensive content in code-mixed and multilingual contexts, the specific problem of identifying racially motivated hoaxes remains insufficiently addressed. Most existing studies focus on broad categories of harmful content, often overlooking the nuanced linguistic and contextual markers unique to racial hoaxes, especially in informal, code-mixed languages like Hinglish. This presents a significant research gap, as racially charged misinformation can have far-reaching societal impacts. In this work, we aim to address this gap by proposing a novel multi-phase fine-tuning approach—first sensitizing models on a related hate speech dataset (THAR), then adapting them to the task-specific HoaxMixPlus dataset for racial hoax detection. This strategy enhances model performance in low-resource settings while introducing a focused lens on racial misinformation.

3 Dataset Description

The datasets used in this work were provided by the organizers of the LT-EDI 2025 shared task [1] (Chakravarthi et al., 2025). Two datasets were employed in our multi-phase fine-tuning approach: the THAR dataset (Sharma et al., 2024), which targets religion-based hate speech, and the HoaxMixPlus dataset, a novel resource annotated for racial hoaxes in code-mixed Hindi-English social media content.

3.1 THAR Dataset

The Targeted Hate Against Religion (THAR) dataset comprises social media comments annotated for the presence of religious hate speech. The dataset consists of binary labels, with values Non-AntiReligion and AntiReligion. This dataset was used to contextually sensitize the model toward sociocultural bias before fine-tuning on the target task. Due to limited data in HoaxMixPlus, we first fine-tune the model on the larger, related THAR dataset to help it learn code-mixed hate speech patterns, enhancing its performance on racial hoax detection.

3.2 HoaxMixPlus Dataset

The HoaxMixPlus dataset consists of 5,105 YouTube comment posts written in code-mixed Hindi-English (Hinglish). It is annotated specifically for racial hoaxes, which are a subcategory of misinformation that falsely associates individuals or groups with crimes or controversial events. This dataset represents an important advancement for low-resource language settings, offering a benchmark for racial hoax detection in multilingual social contexts. The dataset(Training and validation) includes two fields: clean_text and labels, and the test set contains three fields: id, clean_text, and labels. The labels are binary, with values non-racial hoax and racial hoax.

The distribution of both datasets is provided in Table 1 and 2.

Table 1: THAR Dataset Distribution for Religious Hate Speech Detection

Dataset	Non-AntiReligion	AntiReligion	Total
THAR	6,095	5,454	11,549

Table 2: HoaxMixPlus Dataset Distribution for Racial Hoax Detection

Dataset	Non-Racial	Racial	Total
HoaxMixPlus (Train)	2,319	741	3,060
HoaxMixPlus (Dev)	774	247	1,021
HoaxMixPlus (Test)	774	247	1,021

4 Methodology

Text classification remains a fundamental task in Natural Language Processing (NLP), particularly when dealing with complex phenomena such as racial hoaxes in multilingual contexts. Our approach addresses the challenge of detecting racial hoaxes in code-mixed Hindi-English social media content through a novel multi-phase sequential fine-tuning architecture using Hing-BERT model.

This paper aims to highlight the importance of detecting racially motivated hoaxes in online discourse and presents a robust methodology that integrates contextual pretraining, class balancing techniques, and model architecture selection tailored for code-mixed inputs.

4.1 Data Preparation and Balancing

The experiment utilizes two distinct datasets: 'Racial Hoaxes dataset', and 'THAR dataset' (Targeted Hate Against Religion dataset). Due to the inherent class imbalance in the racial hoaxes dataset, we implemented an upsampling technique for the minority class (racial hoaxes) to create a balanced training dataset. This process involves randomly

[1] https://codalab.lisn.upsaclay.fr/competitions/21885

sampling with replacement from the minority class until it matches the size of the majority class, followed by shuffling the combined dataset. This approach prevents bias toward the majority class and improves model generalization.

4.2 Model Architecture

Our approach leverages the "l3cube-pune/Hing-BERT" pre-trained model, which is specifically designed for Hindi-English code-mixed text. This model builds upon the BERT architecture but has been pre-trained on a corpus of code-mixed Hindi-English data, making it particularly suitable for our task. We adapted this model for sequence classification with a binary output layer to classify text as either containing racial hoaxes (1) or not (0). The Hing-BERT model maintains the transformer-based architecture with multiple self-attention heads, which allows it to effectively capture contextual relationships in code-mixed text where linguistic patterns differ significantly from monolingual content.

4.3 Multi-Phase Sequential Fine-tuning

The core innovation in our methodology is the multi-phase sequential fine-tuning approach:

1. **First Fine-tuning Phase (Domain Adaptation/Sensitivity Conditioning):** We initially fine-tune the Hing-BERT model on the THAR dataset focused on anti-religious hate speech content. This phase sensitizes the model's weights towards recognizing nuanced and sensitive linguistic cues commonly present in harmful content. The model thereby develops a refined sensitivity to contextually offensive and hate-indicative language patterns, effectively conditioning it for task adaptation in subsequent fine-tuning stages.

2. **Second Fine-tuning Phase (Task Adaptation):** Building on the sensitized weights from the first phase, we perform a second round of fine-tuning using the racial hoaxes dataset. This phase involves relatively minor adjustments to the preconditioned weights, steering them toward the specific task of detecting racially motivated hoaxes while preserving the model's learned sensitivity to harmful content. This sequential approach allows the model to build upon the knowledge acquired in the domain adaptation phase while specializing in the specific characteristics of racial hoaxes.

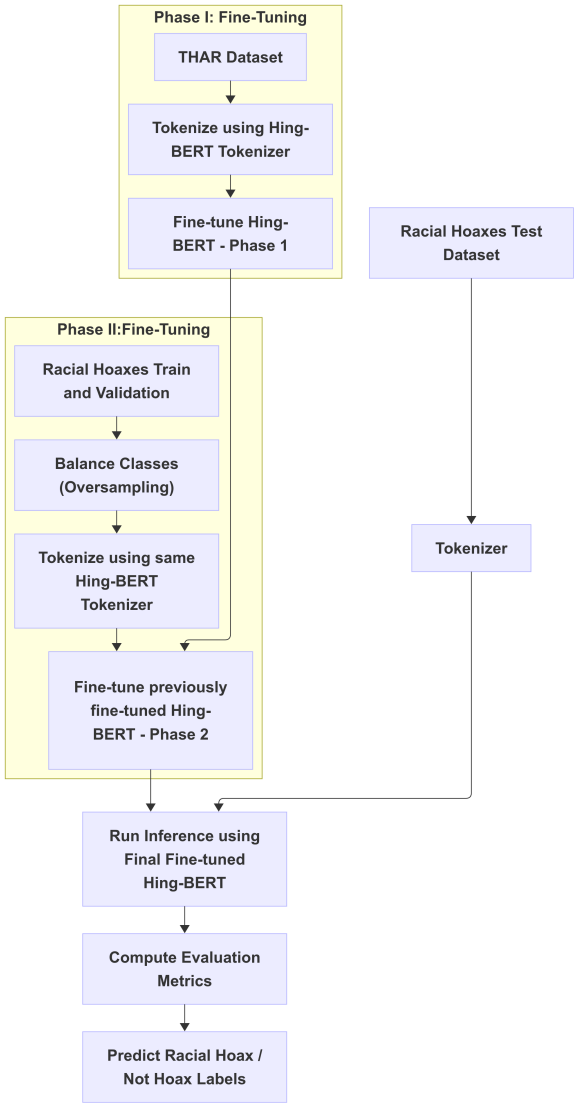

Figure 1: Two-stage fine-tuning and evaluation using THAR and HoaxMixPlus datasets.

This multi-phase approach follows the principle of curriculum learning (Bengio et al., 2009), where the model progressively learns from a broader or similar related domain (religious hate speech) to the specific target domain (racial hoaxes).

4.4 Tokenization and Model Configuration

We employed the specialized tokenizer from "l3cube-pune/Hing-BERT", which effectively handles code-mixed Hindi-English text. Texts were tokenized with a maximum sequence length of 128 tokens, applying padding and truncation as needed. This configuration balances computational efficiency with the need to capture sufficient context from social media posts.

4.5 Training Configuration

The model fine-tuning involves adjusting several standard hyperparameters during the training process, which are set explicitly in the `TrainingArguments` objects. The learning rate is tuned in two phases: Phase I (THAR dataset pre-training) uses a learning rate of 2×10^{-5}, while Phase II (racial hoax dataset fine-tuning) uses a reduced learning rate of 2×10^{-6}. This staged reduction allows the model to converge smoothly and helps prevent catastrophic forgetting after initial domain adaptation. The batch size is set to 16 per device for both training and evaluation. The number of training epochs is set to 4 in both phases. A weight decay of 0.01 is applied to regularize the model and reduce the risk of overfitting. The model saving strategy includes `load_best_model_at_end=True`, ensuring that the best model, based on validation F1-score, is retained at the end of training. All fine-tuning is done by updating the standard transformer layers and the classification head parameters with no weights are frozen. No adapters are used in this model and the fine-tuning directly optimizes the full model without incorporating any additional adapter modules. The code uses `AutoModelForSequenceClassification`, which internally applies cross-entropy loss for binary classification. This is standard and not overridden or custom-defined in the code. The optimization was performed with the AdamW optimizer.

4.6 Inference Pipeline

For deployment and testing, we developed a prediction function that processes new text samples through the following steps:

1. Tokenization of the input text
2. Forward pass through the model
3. Classification based on the output logits
4. Return of human-readable prediction (Racial Hoax detected/Not a Racial Hoax)

Several transformer models were trained using the training and development datasets, and their performance is shown in Table 3. After testing the performance of various transformer models, the top three models with the best performance were selected. The models, Hing-BERT (Nayak and Joshi, 2022), BAAI BGE-M3 (Sun et al., 2024), hing-roberta-mixed (Nayak and Joshi, 2022), MuRIL (Khanuja et al., 2021) were selected. The chosen models were trained using a combined version of the training and validation datasets to make the final predictions.

5 Results

The proposed approach, centered around Hing-BERT and refined using a multi-phase fine-tuning strategy, demonstrated strong effectiveness in identifying racial hoaxes in code-mixed Hindi-English (Hinglish) social media data. After initially adapting the model to socio-religious hate contexts using the THAR dataset, the system was further fine-tuned on the HoaxMixPlus dataset, allowing it to capture task-specific linguistic and contextual cues.

On the test set consisting of 1,021 instances, the model achieved a high overall accuracy of 80%, affirming the robustness of the learned representations. Notably, the model attained a macro-averaged F1-score of 0.73, indicating balanced performance across both hoax and non-hoax classes. The weighted average precision and recall values, both reaching 0.81 and 0.80 respectively, highlight the model's strong capability to make reliable predictions while handling class distribution effectively.

The BAAI BGE-M3 model also performed well, with a macro F1-score of 0.72 and accuracy of 0.79, closely followed by MuRIL and hing-roberta-mixed. Models like Indic-BERT, Hing-BERT LID (Nayak and Joshi, 2022) and roberta-en-hi-codemixed exhibited comparatively lower scores, suggesting they are less effective for this specific task. The results affirm that domain-specific pre-training and code-mixed adaptability significantly enhance model effectiveness for this challenge. The results of all the models evaluated using the training data are presented in Table 3.

The results presented in Table 4 show the per-class F1-scores, precision, and recall for various models in detecting non-racial hoax (Class 0) and racial hoax (Class 1) content.

6 Discussion

Hing-BERT outperforms other models in detecting racial hoaxes within code-mixed Hindi-English social media text due to its alignment with the linguistic characteristics of such data. Unlike models like Indic-BERT, mBERT, or MuRIL, which

Table 3: Model Performance with Multi-Phase Fine-Tuning

Model	Macro F1 Score	Accuracy
Indic-BERT	0.68	0.74
roberta-en-hi-codemixed model	0.67	0.73
BAAI BGE-M3	0.72	0.79
Muril model	0.70	0.75
hing-roberta-mixed	0.70	0.75
Hing-BERT LID	0.69	0.78
Hing-BERT	0.73	0.80

Table 4: Performance Metrics for Each Model (Per-Class F1 Scores)

Model	Non-Hoax			Hoax		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Indic-BERT	0.80	0.87	0.83	0.62	0.50	0.55
roberta-en-hi-codemixed	0.75	0.87	0.81	0.66	0.46	0.54
BAAI BGE-M3	0.87	0.84	0.85	0.56	0.59	0.57
Muril	0.83	0.86	0.85	0.59	0.53	0.56
hing-roberta-mixed	0.76	0.89	0.82	0.70	0.49	0.57
Hing-BERT LID	0.77	0.88	0.69	0.67	0.49	0.56
Hing-BERT	**0.88**	**0.86**	**0.87**	**0.57**	**0.60**	**0.58**

were pre-trained on formal or monolingual corpora, Hing-BERT was pre-trained on large-scale real-world code-mixed data from platforms such as Twitter and YouTube. This exposure enables it to model code-switching patterns, transliteration variants (e.g., *acha*, *accha*, *achha*), and the blending of grammatical structures across languages more effectively.A key strength is its ability to deal with the informal, messy nature of social media, including slang, spelling variations, hashtags, emojis, and subtle code-switch points that may signal sarcasm or misinformation. However, the model has its drawbacks. It tends to favor the majority class due to dataset imbalance and can be sensitive to noisy inputs like excessive emojis or special characters. There's also the risk of hidden biases linked to demographics or dialects in the training data. Finally, the model's decisions are not easily explainable, making it harder to understand or trust why certain posts are flagged as racial hoaxes.

7 Conclusion and Future Work

This study presents an effective multi-phase fine-tuning approach for detecting racial hoaxes in Hinglish social media content, achieving a macro F1-score of 73% using Hing-BERT. By incorporating bias-aware pretraining via the THAR dataset and addressing class imbalance through strategic sampling, the model demonstrates enhanced contextual sensitivity and robustness. Future enhancements may include using contrastive learning (Chen et al., 2020) to better identify subtle forms of hate, incorporating other types of information such as images and hashtags, and expanding the system to support more code-mixed languages spoken in India and fairness audits and bias mitigation to improve reliability. Techniques like adversarial testing (Goodfellow et al., 2015) and explainability (Ribeiro et al., 2016; Lundberg and Lee, 2017) can also help make the model more reliable and easier to understand when used in real-world settings.

8 Source code availability

https://github.com/Abhi-3022/Detecting-Racial-Hoaxes-in-Code-Mixed-Hindi-English-Social-Media-Data

References

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code-mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48. ACM.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR 2015*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Divyanshu Kakwani, Vedanuj Goswami, Janani Prabhakar, Anoop Kunchukuttan, and Pratyush Kumar. 2020. Indiccorp and muril: Large-scale language models for indian languages. *arXiv preprint arXiv:2005.00085*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3406–3412.

Raghav Kumar, Kunal Sinha, Saurabh Varshney, and Manish Shrivastava. 2020. Hingbert: A hinglish language model. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 47–51.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 774–790, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Heng Rathpisey and Teguh Bharata Adji. 2022. Handling imbalance issue in hate speech classification using sampling-based methods. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2583–2592, Gyeongju, Republic of Korea. IEEE.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

K. Shanmugavadivel, A. Narayanan, and M. Senthilkumar. 2022. The challenge of detecting online hate speech: A systematic review. *International Journal of Computer Applications*, 184(12):1–8.

D. Sharma, V. Gupta, V. K. Singh, and B. R. Chakravarthi. 2025a. Stop the hate, spread the hope: An ensemble model for hope speech detection in english and dravidian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

D. Sharma, T. Nath, V. Gupta, and V. K. Singh. 2025b. Hate speech detection research in south asian languages: A survey of tasks, datasets and methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(3):1–44.

Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2024. Thar: Targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

A. Singh, D. Sharma, and V. K. Singh. 2025. Misogynistic attitude detection in youtube comments and replies: A high-quality dataset and algorithmic models. *Computer Speech & Language*, 89:101682.

Xiaofei Sun, Xu Han, Hao Sun, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. Bge-m3: A foundational model for multilingual, multimodal, and multitask retrieval. *Preprint*, arXiv:2403.17818.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of the Association for Information Science and Technology*, 71(12):1475–1487.

CVF-NITT@LT-EDI-2025: A Vision-Language Approach for Detecting Misogynistic Memes in Chinese Social Media

Radhika K T, Sitara K
National Institute of Technology, Tiruchirappalli, India
406322003@nitt.edu,sitara@nitt.edu

Abstract

Online platforms have enabled users to create and share multimodal content, fostering new forms of personal expression and cultural interaction. Among these, memes—combinations of images and text—have become a prevalent mode of digital communication, often used for humor, satire, or social commentary. However, memes can also serve as vehicles for spreading misogynistic messages, reinforcing harmful gender stereotypes, and targeting individuals based on gender. In this work, we investigate the effectiveness of various multimodal models for detecting misogynistic content in memes. We propose a BERT+CLIP+LR model that integrates BERT's deep contextual language understanding with CLIP's powerful visual encoder, followed by Logistic Regression for classification. This approach leverages complementary strengths of vision-language models for robust cross-modal representation. We compare our proposed model with several baselines, including the original CLIP+LR, and traditional early fusion methods such as BERT + ResNet50 and CNN + InceptionV3. Our focus is on accurately identifying misogynistic content in Chinese memes, with careful attention to the interplay between visual elements and textual cues. Experimental results show that the BERT+CLIP+LR model achieves a macro F1 score of 0.87, highlighting the effectiveness of vision-language models in addressing harmful content on social media platforms.

1 Introduction

Misogyny is broadly defined as the hatred of, aversion to, or prejudice against women. It manifests in various forms, including verbal abuse, stereotyping, objectification, or the dissemination of harmful content through social media. Misogyny often appears subtly or overtly in multimodal formats, combining text and imagery to demean, ridicule, or marginalize women. Online misogyny represents a pervasive and deeply rooted societal issue that perpetuates gender-based discrimination and inequality in virtual environments. This toxic behavior not only undermines efforts toward achieving gender equity but also significantly discourages women's active participation in online platforms—ranging from social media to professional and educational digital spaces—thereby silencing their voices and limiting their opportunities for expression, representation, and empowerment(Mohasseb et al., 2025). The increasing prevalence of such content on social media platforms necessitates robust detection systems to mitigate its harmful societal impact.

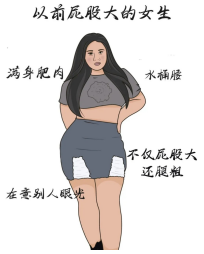

Figure 1: Misogynistic Meme

The Shared Task on Misogyny Meme Detection at LT-EDI@LDK 2025 focuses on building automatic systems to classify memes as *misogynistic* or *non-misogynistic*. This task is especially complex due to the combination of vision and language, as well as the multilingual nature of social discourse—this edition emphasizes Chinese language content. The task encourages advancements in multimodal classification and promotes responsible AI development.

Figure1 depicts a misogynistic meme, shows a cartoon-style illustration of a woman with various derogatory labels surrounding her body. The central theme of the text is body shaming, targeting women with larger body types. The top caption refers to "girls who used to have big butts", setting a critical tone from the start. Additional phrases placed around the figure describe her as "covered

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 47–53

in fat", having a "barrel waist", and "not only a big butt, but also thick legs". Another phrase suggests that she "cares about others' opinions", implying insecurity or social pressure. Together, these captions convey a negative and mocking portrayal of women who do not conform to conventional beauty standards, specifically critiquing body size and shape. This image exemplifies misogynistic and fatphobic content, as it reinforces harmful stereotypes and societal expectations about women's appearance.

The field of image classification has witnessed considerable advancements due to deep learning models like Convolutional Neural Networks (CNN)(Kalchbrenner et al., 2014) and transformers(Kalyan et al., 2022), which have revolutionized the way we process and understand visual data. In parallel to these advancements, Large Language Models (LLMs), have transformed natural language processing by enabling more nuanced understanding and generation of text(Naveed et al., 2023). The integration of these two fields—vision and language—has led to the development of Vision-Language Models (VLM), which combine the strengths of both visual and textual data. These models often use both an image encoder and a text encoder to generate embeddings, which can be fused for various multimodal tasks(Ghosh et al., 2024).

In this work, we utilize the vision-language model Contrastive Language–Image Pre-training (CLIP), introduced by OpenAI, to detect misogynistic content in Chinese memes. CLIP is a powerful pretrained model that maps images and text into a shared embedding space, enabling effective joint understanding of visual and textual modalities. We explore CLIP's capabilities for image representation and pair it with a traditional Logistic Regression (LR) classifier, striking a balance between performance and computational efficiency. We conducted experiments with traditional multimodal baselines, including Bidirectional Encoder Representations from Transformers (BERT) (Vaswani et al., 2017) combined with Residual Networks (ResNet50) (He et al., 2016), as well as Convolutional Neural Networks (CNN) with InceptionV3 (Szegedy et al., 2016), to benchmark performance across different model architectures. BERT+CLIP+LR model performed well among the models and is made available as an open-source resource on GitHub[1].

2 Related Works

The study by (Lei et al., 2024), presents an explainable hateful meme detection model that employs uncertainty-aware dynamic fusion to improve both generalization and interpretability. By dynamically evaluating the uncertainty of visual and textual modalities, the model assigns adaptive weights for feature fusion. They report that visual features are more influential than textual ones in hateful meme detection, and the model's interpretability aids in understanding its decision-making process, although fairness remains a concern for future work.

The work by (Rizzi et al., 2024) proposes a probabilistic framework for detecting elements of disagreement in misogynistic memes by analyzing both the visual and textual components. It explores various strategies for leveraging these elements to identify instances where annotators may disagree in their interpretations. The EXIST 2024 shared task (Vetagiri et al., 2024) focuses on advancing research in detecting and countering sexism on social networks, a persistent and complex societal issue. CNN-BiLSTM for text and ResNet50-CNN-BiLSTM for memes was presented in the paper(Vetagiri et al., 2024) to better identify explicit and implicit sexist content. The task fosters the development of effective strategies through a competitive framework aimed at improving content moderation.

Study by (Ramamoorthy et al., 2022) marked a significant step forward in understanding memes by creating carefully labeled, high-quality data for analyzing sentiment, classifying emotions, and gauging their intensity. To demonstrate the value of this resource, they established initial performance benchmarks using both a text-based model and a multimodal model that integrated visual features with textual understanding. Their findings highlighted the advantage of considering both text and image content for achieving better results in various meme analysis tasks.

(Ponnusamy et al., 2024) introduced the Misogyny Detection Meme Dataset(MDMD), an annotated resource focused on online misogyny within Tamil and Malayalam-speaking communities, offering valuable insights into gender bias and supporting

[1] https://github.com/CyMa-AI/CVF-NITT-LDK2025.git

efforts to combat digital gender-based discrimination. Additionally, the literature review highlights that the Shared Task on Misogyny Meme Detection at LT-EDI@LDK 2025 marks the focused initiatives aimed at detecting misogyny in memes.

3 Methodology

Vision-Language Models can be broadly categorized into three types: (1) Vision-Language Understanding (VLU) models, which interpret and reason over visual and textual inputs; (2) Text Generation with Multimodal Input, where models generate coherent text based on both image and text inputs; and (3) Multimodal Input-Output models, capable of processing and generating across multiple modalities. CLIP belongs to the VLU category, as it learns joint image-text representations through contrastive pretraining (Li et al., 2023). We build upon CLIP by proposing a BERT+CLIP+LR model, where BERT replaces CLIP's text encoder to enhance contextual language understanding. The resulting image and text embeddings are fused and classified using Logistic Regression. To benchmark performance, we also evaluate the original CLIP+LR model and traditional early fusion baselines such as BERT + ResNet50 and CNN + InceptionV3, enabling a comparative analysis of modern VLMs versus conventional multimodal approaches for harmful content detection.

3.1 Problem Definition

Let $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ represent a dataset of memes, where each data sample is a pair (x_i, y_i) with $x_i \in \mathcal{X}$ denoting a meme and $y_i \in \mathcal{Y}$ indicating whether the meme contains misogynistic content ($y_i = 1$) or not ($y_i = 0$).
Each meme x_i consists of two modalities: an image component v_i and a text component t_i, so that $\mathcal{X} = (\mathcal{V}, \mathcal{T})$. The task of misogynistic meme detection is formulated as a binary classification problem. The goal is to learn a predictive function:

$$f : \mathcal{V} \times \mathcal{T} \rightarrow \mathcal{Y}$$

which determines whether a given meme $x_i = (v_i, t_i)$ expresses misogynistic content.

3.2 Data preprocessing

Image part of meme is loaded and preprocessed through a pipeline that includes resizing, center cropping, normalization using predefined mean and standard deviation values, and conversion to a tensor. Faulty, corrupted, or unreadable image files are either skipped or replaced with zero or NaN vectors to maintain consistent batch dimensions and prevent downstream errors during model inference. For the textual modality, each caption or transcription is tokenized, lowercased, and then either truncated or padded to fit the model's maximum token length. Missing or malformed text inputs are handled by substituting neutral placeholders such as "[UNK]" tokens or blank vectors, ensuring input consistency across the dataset.

3.3 Feature Extraction in the Proposed Models

1. **CLIP+LR** :
   The general architecture of the CLIP-based classification model involves separate image and text encoders that generate embeddings, which are then fused into a unified feature vector. CLIP jointly embeds images and text into a shared 512-dimensional space. LR, a widely used linear classification algorithm, is employed for its simplicity, interpretability, and efficiency in high-dimensional spaces (Hosmer Jr et al., 2013). The fused vector obtained from CLIP is passed to the LR classifier to predict the final class label based on the combined visual and textual information. The overall process is illustrated in Figure 2.

2. **BERT+CLIP+LR**: The general architecture of the BERT+CLIP+LR classification model builds upon CLIP by replacing CLIP's original text encoder with BERT(Vaswani et al., 2017), a powerful transformer-based language model known for its deep contextual understanding. In this setup, image inputs are encoded using CLIP's visual encoder, while textual inputs are processed using BERT. The resulting image and text embeddings are concatenated or fused into a single feature vector representing both modalities. This fused vector, which captures complementary visual and linguistic information, is then fed into a LR classifier.

3. **BERT+ResNet50**: ResNet (He et al., 2016) is a deep convolutional neural network known for its ability to train very deep networks using residual connections, making it highly effective for image classification tasks. In

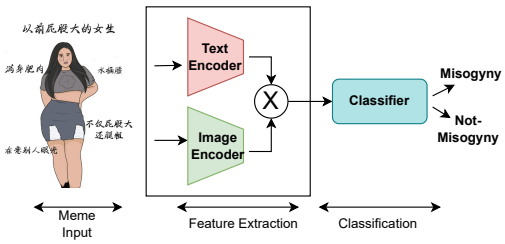

Figure 2: General work flow of proposed models

our multimodal setup, textual features are extracted using BERT, while visual features are obtained from ResNet50. These feature vectors are then concatenated and passed to a classification layer, enabling the model to jointly reason over both modalities. This fusion allows the model to detect nuanced cases of misogyny that may arise from the interaction between image and text in memes.

4. **CNN+InceptionV3**: CNNs treat text as a sequence of word embeddings and learn to capture local dependencies and hierarchical features within the text. For the image component, we employ InceptionV3 (Szegedy et al., 2016), an advanced CNN architecture that uses a multi-branch design to capture features at various levels of abstraction. The multi-branch design enables InceptionV3 to efficiently process images by learning both fine-grained details such as edges and textures and larger, high-level patterns.

3.4 Early Fusion

Instead of explicit concatenation, CLIP encodes each modality independently and enables implicit early fusion through its contrastive pretraining objective. By aligning image and text embeddings in a shared semantic space, CLIP naturally captures cross-modal relationships without requiring manual fusion strategies. In contrast, for traditional models, we implemented explicit early fusion by first extracting textual features using CNN or BERT and visual features using InceptionV3 or ResNet50. These unimodal embeddings were then concatenated to form a joint multimodal feature vector, which was fed into downstream classification. This approach, while effective in controlled settings, often lacks the semantic alignment benefits of contrastively pretrained models.

3.5 Classification

In the proposed model, BERT+CLIP+LR, we employed a logistic regression classifier trained on the aligned image and text embeddings produced by CLIP's encoders. This lightweight classifier proved effective in leveraging CLIP's pretrained semantic representations for binary meme classification, Misogyny vs. Not Misogyny. The final layer outputted probability scores used to determine the predicted class. In the traditional models,Using CNN, BERT, InceptionV3, or ResNet50 for feature extraction, the concatenated multimodal vectors were passed to a dense neural network or a fully connected classification layer.

4 Experiment setup

This section presents the dataset details and describes the experimental configuration used to train and evaluate our models.

4.1 Dataset Description

The dataset developed by (Ponnusamy et al., 2024), provided for the training and development phase, contains the file name of each meme image along with its associated transcribed text. Each meme is annotated with a binary label indicating whether it is misogynistic or non-misogynistic. This structured format allows to develop models that can analyze both visual and textual components of memes. Given the presence of Chinese-language text and complex visual cues, the dataset poses a multilingual and multimodal challenge, encouraging the use of advanced techniques in vision-language understanding for accurate classification. The dataset details are given in Table1.
The dataset is partitioned into 1190 training samples, 170 development/validation samples, and 340 test samples sets, following a roughly 70:10:20 ratio to support robust training, hyperparameter tuning, and final evaluation. The Misogyny class contains 349 training, 47 validation, and 104 test samples, while the Not-Misogyny class includes 841 training, 123 validation, and 236 test samples. The same preprocessing steps are uniformly applied to all three subsets to maintain input consistency during training and evaluation phases.

4.2 Hyperparameters and Model Configuration

In our experiments, we utilized four different models for multimodal meme classification:

Class	Train	Dev	Test
Misogyny	349	47	104
Not-Misogyny	841	123	236
Total	1190	170	340

Table 1: Dataset details

BERT+CLIP+LR, CLIP+LR, BERT+ResNet50, and CNN+InceptionV3, each with distinct hyperparameters.

For CLIP+LR, the image and text features were extracted using CLIP's pretrained Vision Transformer (ViT-B-32) for text and the corresponding image encoder. Both image and text embeddings were normalized using L2 normalization, and logistic regression was employed as the classifier, with a learning rate of 2e-5 for feature extraction and training.

In the BERT+ResNet50 model, BERT was used for text feature extraction, and ResNet50 was employed for image features. Both models were fine-tuned using the Adam optimizer with a learning rate of 2e-5, and the CrossEntropy Loss was used for classification.

In the CNN+InceptionV3 setup, images were processed using a pre-trained InceptionV3 model, without the top classification layer, with the input image size set to (299, 299, 3). The output from the InceptionV3 was passed through a GlobalAveragePooling layer and a fully connected layer with 128 units and ReLU activation. The model was compiled with the Adam optimizer, using a learning rate of 1e-4 and categorical cross-entropy loss for classification.

5 Experimental Evaluation

We evaluated our multimodal architectures for misogyny meme classification using standard metrics, including accuracy, macro precision, macro recall, and macro F1-score.

5.1 Overall Performance

The BERT+CLIP+LR architecture achieved the best overall performance in our experiments. By combining CLIP's powerful pretrained visual encoder with BERT's deep contextual text representations, the model captured cross-modal relationships more effectively than the original CLIP+LR setup or conventional fusion strategies. This approach enhanced textual understanding while retaining the efficiency and semantic alignment benefits of CLIP's visual features. Despite the architectural simplicity—without relying on complex deep fusion layers—BERT+CLIP+LR delivered superior accuracy while remaining computationally efficient and interpretable.

Model	Precision	Recall	Macro F1	Weighted F1	Accuracy
CNN+InceptionV3	0.80	0.81	0.76	0.80	0.81
BERT+ResNet50	0.85	0.84	0.82	0.85	0.84
CLIP+LR	0.86	0.86	0.83	0.86	0.86
BERT+CLIP Image Encoder+LR	**0.89**	**0.89**	**0.87**	**0.89**	**0.89**

Table 2: Evaluation of proposed models for misogynistic meme detection

Model	Total Inference Time (s)	Avg. Time per Sample (s)
CNN + InceptionV3	8.0840	0.0238
CLIP + LR	20.7715	0.0611
BERT + ResNet50	28.2900	0.0832
BERT + CLIP Image Encoder + LR	25.2456	0.0742

Table 3: Inference time comparison of proposed models.

As shown in Figure 3, the BERT+CLIP+LR model correctly classified 214 non-misogynistic and 87 misogynistic samples, with 22 false positives and 17 false negatives. These misclassifications likely stem from class imbalance and limited variability in the training data, which can hinder the model's ability to generalize to ambiguous or nuanced cases.

The early fusion approaches, which combined textual and visual features through concatenation, showed competitive results. Specifically, the combination of BERT for text and ResNET50 for images consistently outperformed CNN-based text and image representations, highlighting the effectiveness of contextual language embeddings in understanding meme text. Summary of classification performance across models are presented in Table2.

We calculated the total inference time and the average inference time per sample for the proposed methods on the test dataset. The results are presented in Table 3. While the CNN + InceptionNet model offers the fastest inference time, CLIP + LR provides a reasonable compromise between performance and inference speed, making it suitable for applications where a balance between accuracy and efficiency is desired.

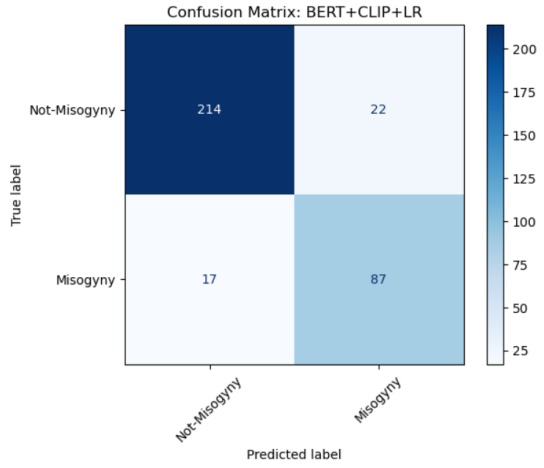

Figure 3: Confusion matrix of multimodal model BERT+CLIP+LR

5.2 Error Analysis

False Positive: Sample278.jpg contains a humorous diagram featuring the Chinese character meaning "woman" at the center of a radar chart, with all attributes of intelligence, courage, perseverance, physical strength, and lifespan. Despite being a positive or humorous depiction celebrating women's qualities, the model incorrectly flagged it as misogynistic. This misclassification likely stems from overfitting to visual or textual stylistic patterns like the presence of gender-related characters or symbols without understanding the broader context or intent. The error highlights the need for improved context-aware classification in multimodal systems.

False Negatives: Sample113.jpg, with text, "Stop talking about that trashy woman" was misclassified as not misogyny despite containing gendered slurs and hostile language toward women. The meme expresses contempt using a derogatory Chinese phrase aimed at women, reinforcing misogynistic sentiment. However, the model likely failed to detect this due to language limitations of non-English text transcription and lack of cultural context, leading to an undetected instance of harmful bias.

6 Conclusion

In this study, we investigated a vision-language multimodal approach for misogynistic meme classification. We began with the CLIP+LR model, where CLIP's contrastive pretraining effectively aligned image and text features without requiring explicit fusion layers, resulting in a model that was both accurate and computationally efficient. Building upon this, we proposed the BERT+CLIP+LR model, which replaces CLIP's text encoder with BERT to capture deeper contextual understanding of language. This enhancement led to improved cross-modal alignment and superior classification performance, while maintaining architectural simplicity. To benchmark our approach, we also implemented traditional early fusion models that combined CNNs and BERT for text with InceptionV3 and ResNet50 for image features. Overall, our findings highlight that Vision-Language Models offer a scalable, robust, and efficient solution for multimodal classification tasks, and represent a promising direction for future research in understanding harmful online content. For future work, we aim to develop misogynistic dataset to include more languages and cultural contexts, explore more advanced vision-language models, and investigate methods to detect implicit and context-dependent misogyny in memes.

References

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Xia Lei, Siqi Wang, Yongkai Fan, and Wenqian Shang. 2024. Hate-udf: Explainable hateful meme detection with uncertainty-aware dynamic fusion. *Software: Practice and Experience*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Alaa Mohasseb, Eslam Amer, Fatima Chiroma, and Alessia Tranchese. 2025. Leveraging advanced nlp techniques and data augmentation to enhance online misogyny detection. *Applied Sciences*, 15(2).

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and 1 others. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: workshop on multimodal fact checking and hate speech detection, CEUR*.

Giulia Rizzi, Paolo Rosso, and Elisabetta Fersini. 2024. From explanation to detection: Multimodal insights into disagreement in misogynous memes.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Advaitha Vetagiri, Prateek Mogha, and Partha Pakray. 2024. Cracking down on digital misogyny with multilate: A multimodal hate detection system. *Working Notes of CLEF*.

Wise@LT-EDI-2025: Combining Classical and Neural Representations with Multi-scale Ensemble Learning for Code-mixed Hate Speech Detection

Ganesh Sundhar S[1], Durai Singh K[1], Gnanasabesan G[1], Hari Krishnan N[1], Dhanush MC[1]
[1]Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22017, cb.en.u4aie22167, cb.en.u4aie22018, cb.en.u4aie22020, cb.en.u4aie22130}
@cb.students.amrita.edu

Abstract

Detecting hate speech targeting caste and migration communities in code-mixed Tamil-English social media content is challenging due to limited resources and socio-cultural complexities. This paper proposes a multi-scale hybrid architecture combining classical and neural representations with hierarchical ensemble learning. We employ advanced preprocessing including transliteration and character repetition removal, then extract features using classical TF-IDF vectors at multiple scales (512, 1024, 2048) processed through linear layers, alongside contextual embeddings from five transformer models-Google BERT, XLM-RoBERTa (Base and Large), SeanBenhur BERT, and IndicBERT. These concatenated representations encode both statistical and contextual information, which are input to multiple ML classification heads (Random Forest, SVM, etc). A three-level hierarchical ensemble strategy combines predictions across classifiers, transformer-TF-IDF combinations, and dimensional scales for enhanced robustness. Our method scored an F1-score of 0.818, ranking 3rd in the LT-EDI-2025 shared task, showing the efficacy of blending classical and neural methods with multi-level ensemble learning for hate speech detection in low-resource languages.

Keywords: Caste/Migration-based hate speech detection, Code-mixed text, Transliteration, TF-IDF Features, Transformer embeddings, Hierarchical ensemble learning, Low-resource languages

1 Introduction

Hate speech is any kind of communication that attacks a person or group based on attributes like caste, religion, race, or other identity factors. With the advancement of technology and the advent of social media, individuals can now share their thoughts with anyone in the world. While this increased connectivity has many benefits, the anonymity offered by online platforms has unfortunately facilitated the spread of hateful messages, particularly those targeting vulnerable groups such as migrants and specific caste communities. To make online communities inclusive, it is essential to identify caste and migration-based hate speech.

Despite significant progress in Natural Language Processing(NLP) through transformer architectures (Vaswani et al., 2017) revolutionizing text classification tasks, detecting hate speech in low-resource languages like Tamil poses unique challenges stemming from dialectal diversity and regional variations. This challenge is further intensified when detecting specific forms of hate speech, such as those targeting caste and migration, as these are often embedded in local socio-cultural nuances and context. In social media platforms, the texts are often code-mixed, i.e., English + Tamil, making detection even more challenging.

To address these challenges, our work presents a multi-scale hybrid architecture that combines classical Term Frequency-Inverse Document Frequency(TF-IDF) (Spärck Jones, 1972) features at multiple scales (512, 1024, and 2048) with contextual transformer embeddings. For final classification, a hierarchical ensemble using majority voting across models and feature scales was used to make the model robust and generalizable.

2 Related Work

Early hate speech detection systems relied on rule-based methods and keyword matching (Clarke et al., 2023). These methods are unsuitable for the vast amount of data present today as they lack contextual understanding. Machine learning algorithms like Support Vector Machines (Kp et al., 2009) and Logistic Regression (Hosmer Jr et al., 2013), which used hand-crafted features such as TF-IDF, n-grams, and Parts of Speech (POS) tags, emerged later. Although these models improved

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 54–62

performance, they were ineffective for code-mixed or culture-specific hate speech (Davidson et al., 2017).

Deep learning techniques such as RNNs (Elman, 1990) and LSTMs (Hochreiter and Schmidhuber, 1997), when combined with word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), improved context-aware modeling and reduced the need for manual feature engineering (Pitsilis et al., 2018). Nevertheless, they struggled with long-range dependencies and noisy, code-mixed text. With the emergence of pre-trained language models such as BERT (Devlin et al., 2019a), RoBERTa (Liu et al., 2019), and multilingual variants like XLM-R (Conneau et al., 2020a), the performance improved further. Fine-tuning these pretrained models on hate speech datasets has consistently yielded better results (Albladi et al., 2025).

In their work, (Roy et al., 2022) used an ensemble-based approach to detect hate speech in code-mixed Tamil and Malayalam texts, as the individual models had a high misclassification rate. Two ensemble techniques were used: one based on the average of the outcomes and another using custom weights. Their ensemble model outperformed the previously reported state-of-the-art models, achieving an F1 score of 0.933 on Tamil and 0.802 for the Malayalam dataset.

In their work, (Sreelakshmi et al., 2024) detected Hate Speech and Offensive Language (HOS) in low-resource Dravidian CodeMix languages (Kannada, Malayalam, Tamil). Various multilingual transformer-based embeddings (e.g., MuRIL, BERT, XLM-R) were combined with traditional ML classifiers for HOS detection. To address class imbalance, a cost-sensitive learning approach was used. Experiments on six datasets showed that MuRIL + SVM performed best overall.

3 Task and Dataset Description

The goal of this shared task (Rajiakodi et al., 2025) is to develop a system to detect hate speech targeting caste and migrant communities in code-mixed social media data, focusing on Tamil, a low-resource language. The dataset (Rajiakodi et al., 2024) has three columns: "text", which has the comments from social media platforms; "id", containing the ID of the comments; and "label", which is set to 1 for hate speech and 0 for non-hate speech. The dataset description is provided in Table 1.

Dataset	No. of comments
Train	5512
Dev	787
Test	1576
Total	7875

Table 1: Distribution of comments across training, development, and test sets

4 Methodology

The approach uses a multi-scale hybrid framework to identify hate speech in code-mixed Tamil social media posts. Two preprocessing schemes (with and without transliteration) are used to handle intrinsic noise in the data, such as emojis, URLs, inconsistent spacing, repeated characters in transliterated Tamil words, and code-mixing between Tamil and English languages. Feature extraction unites TF-IDF vectors at three dimensions (512, 1024, 2048) with contextual embeddings of five transformer models. Such features are then concatenated to create integrated feature vectors with both statistical and contextual information. A set of 22 traditional ML classifiers are trained for each feature set, and the top 3 models for each were chosen. A three-level hierarchical ensemble approach employs majority voting across classifiers, feature sets, and dimensions to ensure resilient classification by combining heterogeneous preprocessing schemes, representation types, and model architectures

4.1 Data Preprocessing

Our data pre-processing pipeline addressed the challenges of social media text containing code-mixed Tamil and English content. In the first approach, without transliteration, newlines were replaced with white spaces, emojis were converted into text i.e. demojization (Kim and Wurster, 2014), URLs were removed, multiple whitespaces were replaced with a single space, and then the text was converted into lowercase.

In the second approach, the same steps were repeated, followed by transliteration (Karimi et al., 2011) of Tamil Unicode characters into their English equivalents, and repeated characters in transliterated Tamil words were removed (while preserving standard English words), and then non-ASCII characters were removed to maintain consistency.

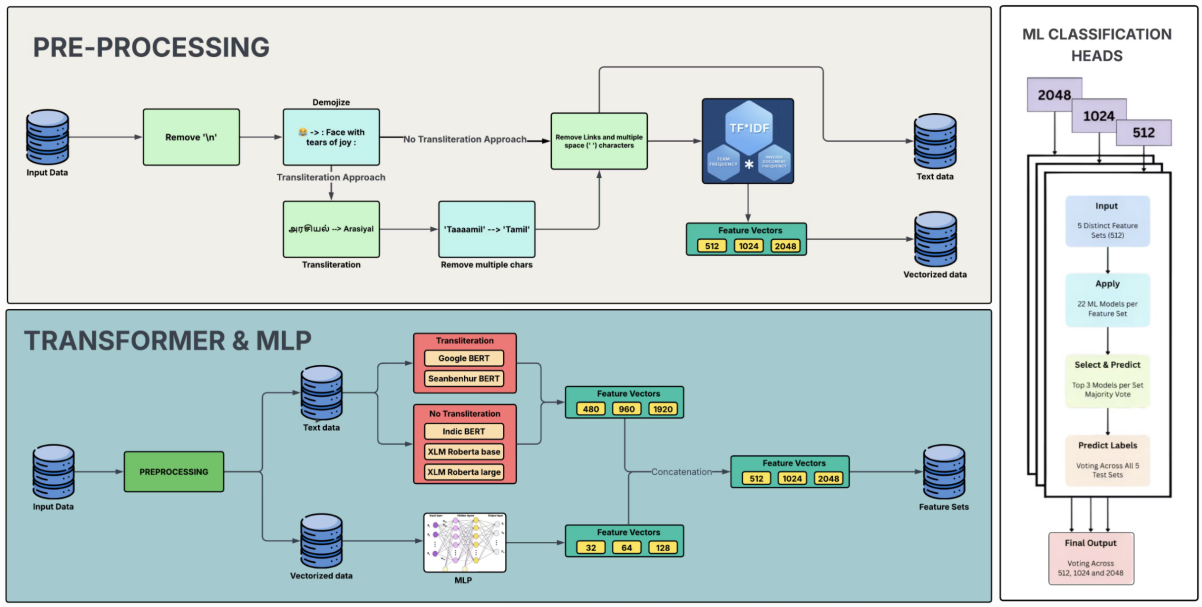

Figure 1: Multi-scale architecture for hate speech detection: Three-stage pipeline showing (a) data preprocessing with dual approaches (with/without transliteration) producing both processed text data and TF-IDF vectors, (b) transformer models generating contextual embeddings that are combined with reduced TF-IDF features to create unified Wise Embeddings (WE), and (c) machine learning classifiers processing WE features followed by three-level hierarchical ensemble voting to produce final hate speech predictions.

4.2 TF-IDF Vectorization

TF-IDF vectorization (S N et al., 2022) was used to extract features from the text, and its hyperparameters were optimized using grid search (Hutter et al., 2019). This resulted in high-dimensional feature vectors consisting of approximately 22,000 features. As these feature vectors are sparse, Truncated Singular Value Decomposition (SVD) was applied:

$$\mathbf{X} \approx \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T \quad (1)$$

This decomposition (Halko et al., 2011) was used to reduce the feature space to three different dimensions: 512, 1024, and 2048. These features were further refined using Feed-Forward Networks (FFNs) (Rumelhart et al., 1986) to produce compact embeddings of 32, 64, and 128 dimensions respectively.

4.3 Transformer Embeddings

Five different transformer models were used to extract contextual embeddings from the preprocessed text. For the models Google BERT (Devlin et al., 2019b) and SeanBenhur BERT (Benhur and Sivanraju, 2021), the input was text preprocessed with transliteration, while for Indic BERT (Kp et al., 2025), XLM RoBERTa base (Conneau et al., 2020b), and XLM RoBERTa large, the input was text preprocessed without transliteration. The contextual representations were extracted from the [CLS] token, which serves as an aggregate representation of the entire input sequence. The transformer models produce embeddings of varying dimensions: IndicBERT, XLM RoBERTa Base, and SeanBenhur BERT generate 768-dimensional embeddings, while Google BERT Large and XLM RoBERTa Large produce 1024-dimensional embeddings. These obtained embeddings undergo linear transformation to achieve target dimensions of 480, 960, and 1920 for the three scales respectively. Each transformed embedding is then concatenated with its corresponding FFN-reduced feature vectors (of dimensions 32, 64, and 128), resulting in unified representations of 512, 1024, and 2048 dimensions that capture both statistical and contextual information. The unified representations obtained are hereafter referred to as Wise Embeddings (WE).

4.4 Machine Learning Models

For each of the five feature sets at every dimension in WE, 22 traditional machine learning classifiers including Logistic Regression, SVM, Naive Bayes (McCallum and Nigam, 1998), and Random Forest Classifier (Breiman, 2001) were trained. From these, the top three classifiers were selected based on their validation performance.

For each feature set and scale, predictions from the top three classifiers were aggregated using majority voting to give a single predicted label. Then, the five resulting predictions for each dimension (5 BERT models) was again combined using majority voting to produce a final label per dimension. Then, a cross-dimensional ensemble was performed, where the three labels i.e. one from each WE dimension underwent another round of majority voting to determine the overall predicted label.

During the initial transformer training phase, the Wise Embeddings (WE) are processed through a final linear layer that maps the 512, 1024, or 2048-dimensional representations to a single output value, optimized using Binary Cross-Entropy (BCE) (Goodfellow et al., 2016) loss. This linear layer effectively functions as a linear classifier, constraining the learned WE representations to be linearly separable in the feature space. However, the complex nature of code-mixed hate speech detection often exhibits non-linear patterns that cannot be adequately captured by linear decision boundaries alone.

By applying traditional ML classifiers with inherently non-linear decision boundaries (such as Random Forest and SVM with non-linear kernels) to these linearly-optimized WE features, we introduce additional modeling capacity to capture complex patterns in the data. This approach leverages the pre-trained linear separability while allowing non-linear classifiers to model intricate relationships that the original linear layer could not capture. The ensemble voting across multiple classifiers further enhances robustness and generalization, particularly beneficial for handling the noisy and heterogeneous nature of code-mixed social media text.

5 Result and Analysis

The model's performance was assessed using the F1-score (Powers, 2011), which is defined in Equation 2.

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C}\frac{2 \times P_i \times R_i}{P_i + R_i} \quad (2)$$

The results of the best-performing models for different dimensions (512, 1024, and 2048) are summarized in Table 2 and the results of different ensembles are given in Table 3. The ensemble of the three different dimensions achieved the highest F1-score of 0.85 in the dev set.

Dim	Transformer	Best Model	F1
512	Google BERT	RF	0.80
	IndicBERT	RF	0.81
	SeanBenhur BERT	XGBoost	0.81
	XLM-R Base	RF	0.81
	XLM-R Large	Extra Trees	0.81
1024	Google BERT	SVM	0.81
	IndicBERT	RF	0.83
	SeanBenhur BERT	RF	0.82
	XLM-R Base	Nu-SVM	0.83
	XLM-R Large	RF	0.83
2048	Google BERT	Gradient Boosting	0.84
	IndicBERT	Ridge Regression	0.82
	SeanBenhur BERT	Nu-SVM	0.83
	XLM-R Base	RF	0.84
	XLM-R Large	RF	0.83

Table 2: PerformancemMetrics of the combinations

Ensemble Type	F1-score
512 dim models	0.81
1024 dim models	0.82
2048 dim models	0.84
Cross-dimensional ensemble	0.85

Table 3: Performance metrics of multi-scale ensembles

5.1 Comparison

Our proposed methodology using Wise Embeddings (WE), achieved an F1-score of 0.81827 on the test set, securing 3rd rank in the shared task. The F1 scores of the top five performing teams in the shared task are summarized in Table 4.

Rank	Team Name	F1-score
1	CUET_N317	0.88105
2	CUET's_white_walkers	0.86289
3	Wise	0.81827
4	CUET_blitz_aces	0.81682
5	hinterwelt	0.80916

Table 4: Top 5 Teams ranked based on F1-score

6 Conclusion

The present work demonstrates the strength of combining neural and conventional representations through a multi-level ensemble approach for caste and migration-based hate speech detection in code-mixed Tamil text. The approach highlights the importance of employing diverse feature representations in addressing challenging NLP problems in low-resource languages.

Github Source code: https://github.com/Ganesh2609/CasteMigrationHateSpeech

7 Limitations

Even though the proposed pipeline performed well, there are a few limitations, which are as follows:

1. The relatively small dataset (7,875 comments in total) may limit the model's ability to generalize across the full spectrum of hate speech variations in Tamil social media content.
2. Data inconsistency exists where identical comments appear in both training and development sets with conflicting labels, potentially compromising the model's learning process and evaluation reliability.
3. Label quality issues are present in the dataset, where some clearly hateful content is labeled as non-hate speech, while certain benign comments are marked as hate speech. This annotation ambiguity, which is challenging even for human annotators, introduces noise that may affect model performance.

References

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Sean Benhur and Kanchana Sivanraju. 2021. Pretrained transformers for offensive language identification in tanglish. *arXiv preprint arXiv:2110.02852*.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint, arXiv:1703.04009*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated machine learning: methods, systems, challenges*. Springer Nature.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Computing Surveys (CSUR)*, 43(3):1–46.

Taehoon Kim and Kevin Wurster. 2014. emoji: Emoji for python. https://github.com/carpedm20/emoji.

Soman Kp, Rajendran Loganathan, and Ajay Vadakkepatt. 2009. *Machine learning with SVM and other kernel methods*.

Suriya Kp, Durai Singh K, Vishal A S, Kishor S, and Sachin Kumar S. 2025. Synapse@DravidianLangTech 2025: Multi-class political sentiment analysis in Tamil X (Twitter) comments: Leveraging feature fusion of IndicBERTv2 and lexical representations. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 716–720, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint, arXiv:1907.11692*.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. AAAI Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint, arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

David Martin Powers. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Prasanth S N, R Aswin Raj, Adhithan P, Premjith B, and Soman Kp. 2022. CEN-Tamil@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using TF-IDF and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74, Dublin, Ireland. Association for Computational Linguistics.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21. Reprinted in Journal of Documentation, Vol. 60, No. 5, pp. 493-502, 2004.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Training Performance Metrics

This appendix presents the training and validation performance metrics for all transformer models across the three dimensional scales (512, 1024, and 2048). Each figure shows the loss, accuracy, and F1-score curves during the training process.

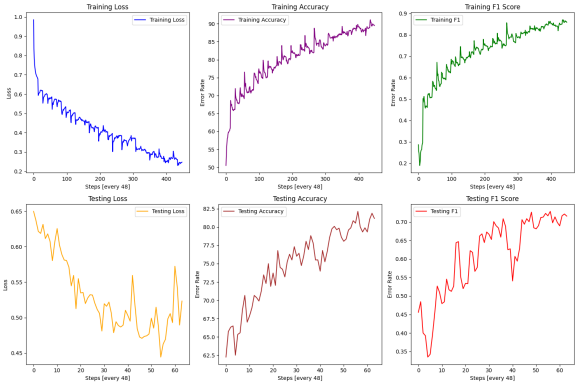

Figure 2: Training and validation metrics for Google BERT with 512-dimensional embeddings.

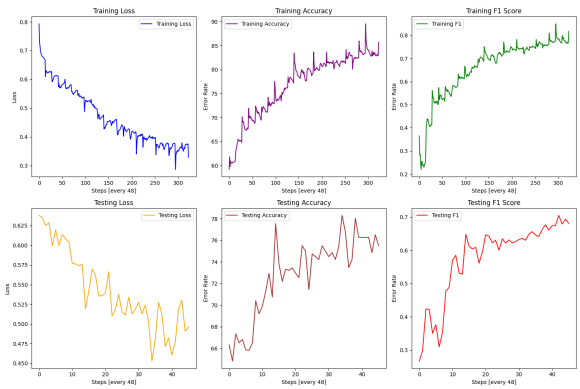

Figure 3: Training and validation metrics for IndicBERT with 512-dimensional embeddings.

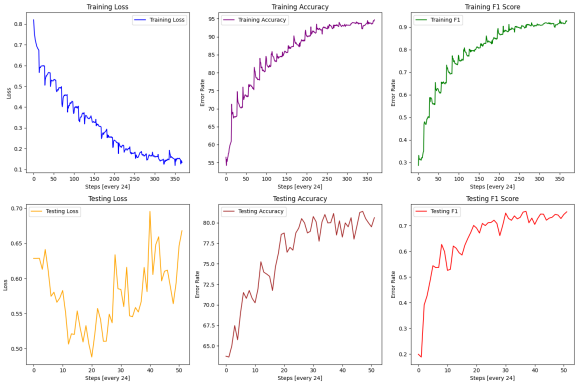

Figure 4: Training and validation metrics for SeanBenhur BERT with 512-dimensional embeddings.

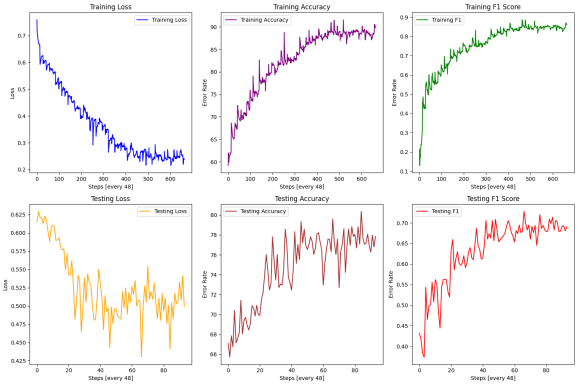

Figure 5: Training and validation metrics for XLM-RoBERTa Base with 512-dimensional embeddings.

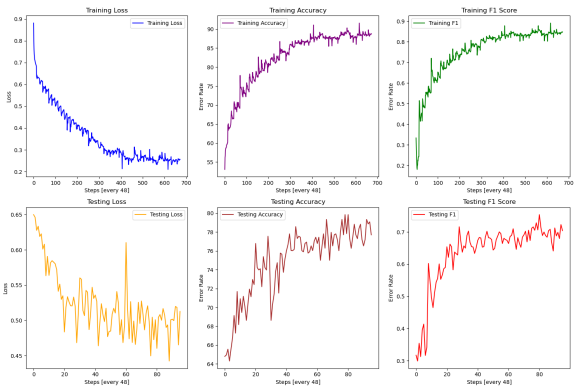

Figure 6: Training and validation metrics for XLM-RoBERTa Large with 512-dimensional embeddings.

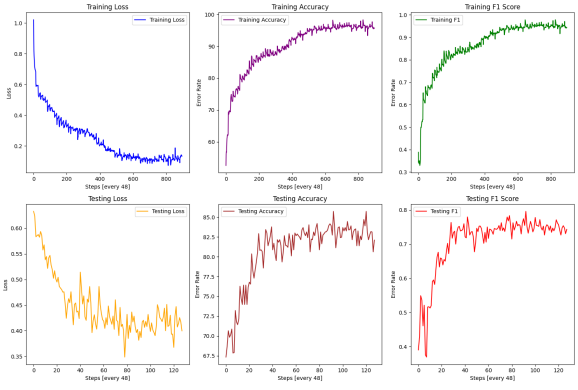

Figure 7: Training and validation metrics for Google BERT with 1024-dimensional embeddings.

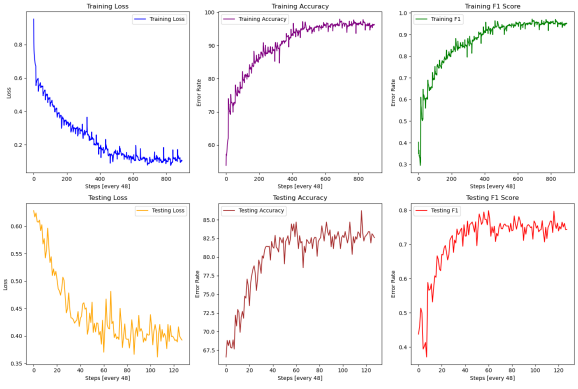

Figure 8: Training and validation metrics for IndicBERT with 1024-dimensional embeddings.

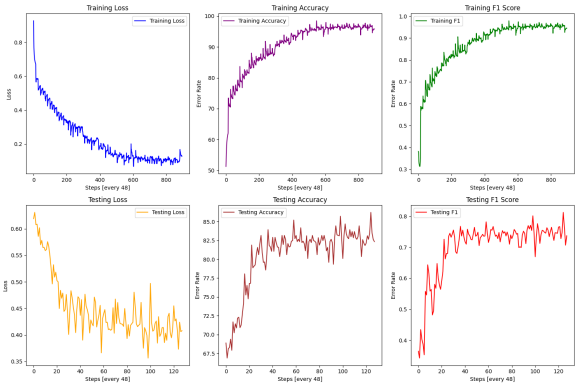

Figure 9: Training and validation metrics for SeanBenhur BERT with 1024-dimensional embeddings.

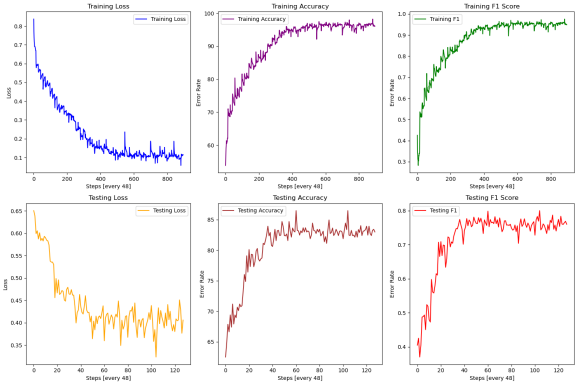

Figure 10: Training and validation metrics for XLM-RoBERTa Base with 1024-dimensional embeddings.

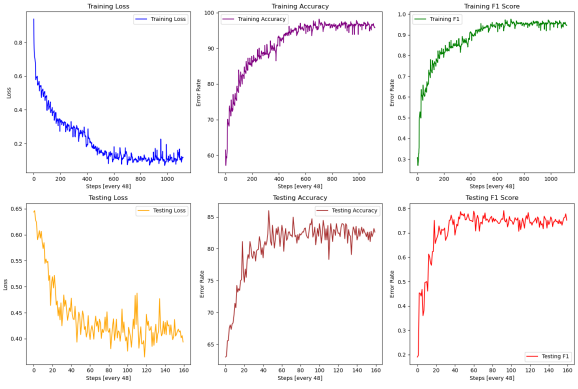

Figure 11: Training and validation metrics for XLM-RoBERTa Large with 1024-dimensional embeddings.

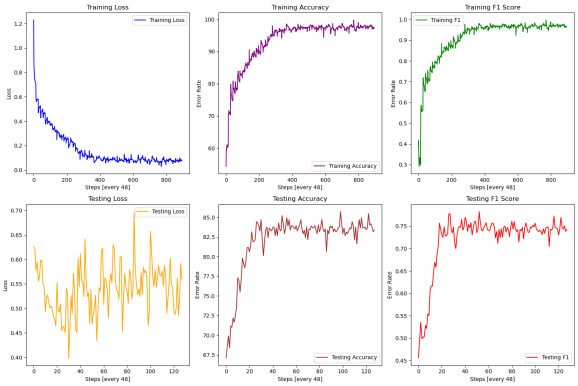

Figure 12: Training and validation metrics for Google BERT with 2048-dimensional embeddings.

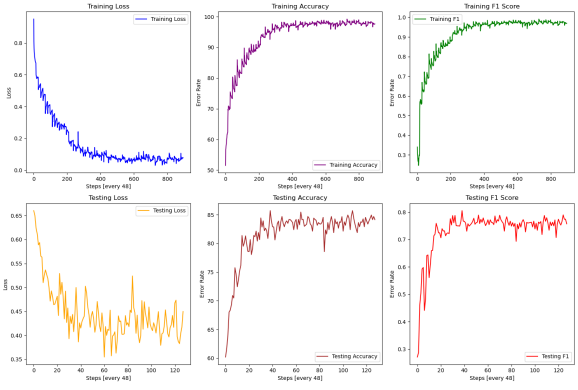

Figure 13: Training and validation metrics for IndicBERT with 2048-dimensional embeddings.

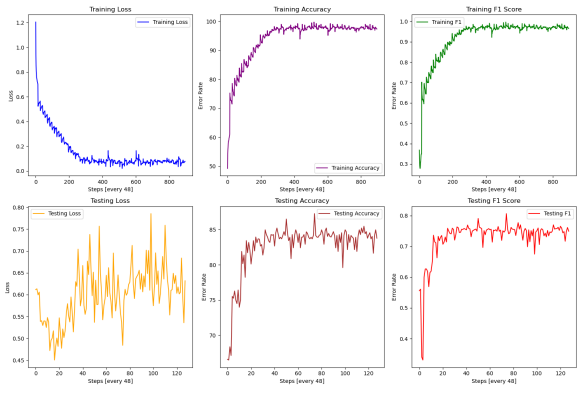

Figure 14: Training and validation metrics for SeanBenhur BERT with 2048-dimensional embeddings.

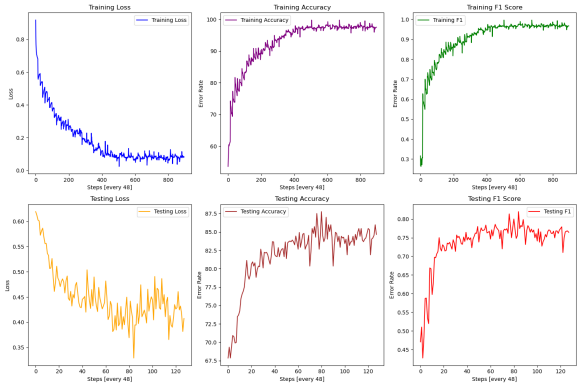

Figure 15: Training and validation metrics for XLM-RoBERTa Base with 2048-dimensional embeddings.

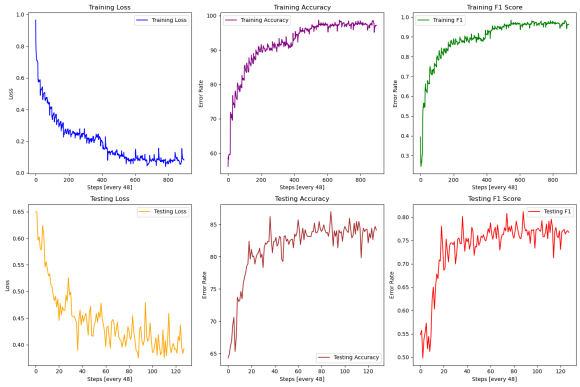

Figure 16: Training and validation metrics for XLM-RoBERTa Large with 2048-dimensional embeddings.

CUET's_White_Walkers@LT-EDI 2025: Racial Hoax Detection in Code-Mixed on Social Media Data

Md Mizanur Rahman, Jidan Al Abrar, Md Siddikul Imam Kawser, Ariful Islam, Md. Mubasshir Naib, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{116, 080, 081, 129, 089}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

False narratives that manipulate racial tensions are increasingly prevalent on social media, often blending languages and cultural references to enhance reach and believability. Among them, racial hoaxes produce unique harm by fabricating events targeting specific communities, social division and fueling misinformation. This paper presents a novel approach to detecting racial hoaxes in code-mixed Hindi-English social media data. Using a carefully constructed training pipeline, we have fine-tuned the XLM-RoBERTa-base multilingual transformer for training the shared task data. Our approach has incorporated task-specific preprocessing, clear methodology, and extensive hyperparameter tuning. After developing our model, we tested and evaluated it on the LT-EDI@LDK 2025 shared task dataset. Our system achieved the highest performance among all the international participants with an F1-score of 0.75, ranking 1st on the official leaderboard.

1 Introduction

Racial hoaxes on social networks have continuously emerged as a significant concern, which may lead to increased ethnic tension and social unrest. In this study, we define Hoax Speech as intentional, deceptive linguistic content that mimics the tone or structure of hate speech or propaganda, yet lacks genuine hateful intent. It often uses irony, satire, or fabricated narratives to incite confusion or mask malicious undertones for a criticise race. Racial hoaxes refer to fabricated statements that falsely accuse specific racial groups with malicious intent. These intents are designed to manipulate public opinion and provoke communal or ethnic tension.

According to Amnesty International in The News Minute, Racial hoaxes and hate speech have become prevalent in South Asia day by day, leading to violence and social turmoil. For instance, between 2015 and 2019, India witnessed 902 alleged hate crimes resulting in 303 deaths, with Muslims constituting the majority of victims [1]. Social media can amplify such crimes, leading to hatred, misinformation for the minority group or community, especially. A significant number of previous studies have been conducted on spreading misinformation, fake news, and hate speech detection (Barker and Jurasz, 2021), (Arellano et al., 2022), but the task of racial hoax detection has not been explored too much.

The primary objective of this paper is to detect Hindi-English code-mixed racially deceptive text on social media platforms. We have used a transformer-based model, HinG-RoBERTa, which is pre-trained on Hindi-English code-mixed data, to train our model. The core contributions of our research work are as follows:-

1. We have developed an effective model to detect racial deception in Hindi-English code-mixed text.
2. We have conducted a series of experiments on the dataset and comprehensively analyzed their performance outcomes.

The implementation details have been provided in the following GitHub repository:- https://github.com/Mizan116/LT-EDI-LDK-2025/Racial Hoax.

2 Related Study

Hate speech detection identifies degrading information, especially on social media. Hate speech, sexism, homophobia, racism, bullying, and other verbal abuse are detected. Prior work has studied the online dissemination of racially based stereotypes

[1] 2019 sees steepest rise in hate crimes since 2016, finds Amnesty tracker

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 63–67

and disinformation. However, very few studies examine racial hoaxes through multilingual lenses (Bourgeade et al., 2023).

Initially, the discipline was dominated by classical machine learning algorithms such as Support Vector Machines (SVMs) and Naïve Bayes and Random Forests for text mining techniques. The authors of Afroz et al. (2012) used machine learning techniques to detect hoaxes in the English language. Rule-based machine learning methods have also been used in Chopra et al. (2020). They have been used for detecting hate speech in Hindi-English code-mixed text. Research across multiple languages shows how racial hoaxes and stereotypes circulate in social media conversations in Bourgeade et al. (2023). In a related shared task on Dravidian languages, the authors of Rahman et al. (2025) employed transformer models like XLM-R and MuRIL, demonstrating high performance in abusive language detection. The authors of Ahmed et al. (2022) have detected hateful users more accurately and fairly, including social network context. On the other hand, (Papapicco et al., 2022) researched how adolescents show confidence in spotting fake news, but often fail to detect or remember racial hoaxes.

According to Ahmed et al. (2022), the integration of social network data contributes to improved performance and equity in classification systems. The dearth of developed critical thinking skills exposes teenagers to greater deception. Adolescents often feel immune to fake news despite being unable to identify or remember it (Papapicco et al., 2022). The Biradar et al. (2024b) introduced a novel dataset with Hindi-English code-mixed dataset of hateful comments, aiming to explore the link between fake narratives and hate severity. It is problematic to identify hate speech in code-switched Languages such as Hinglish because of its intricacy. Some social network analyses have applied a focus on features of social media such as usual name-calling but study of bias elimination and diversity linguistics is scant (Chopra et al., 2020).

3 Dataset & Task Overview

We have utilized the abusive detection dataset from the LT-EDI@LDK 2025 shared task (Chakravarthi et al., 2025). This research makes use of a code-mixed Hindi-English dataset meant for detecting racial hoaxes in social media posts. The dataset is publicly available in Chakravarthi (2020) research. We used their dataset to train, test and evaluate our model. The dataset is divided into three parts: training, development, and test. The test set does not have labels. Our models predicted the labels for that set. The dataset has two 'Non-Hoax' (0) and 'Racial Hoax' (1) annotations per sample. We have training (3060 samples), validation (1021 samples), and test (1021 samples). Around each text sample, there is a word count of 29-30. The dataset suffers from class imbalance (76% - 24%) as the Non-Hoax class dominates strongly over the Racial Hoax class.

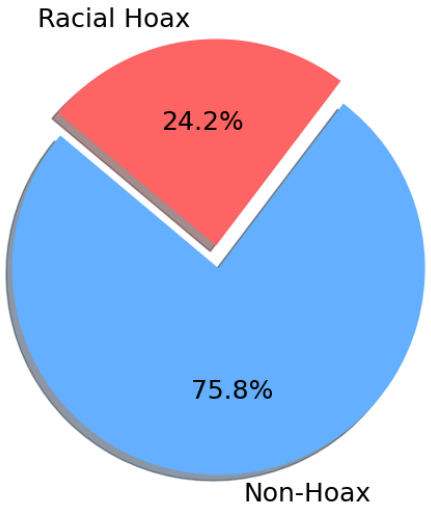

Figure 1: Training Data Distribution

Split	Non-Hoax (0)	Racial Hoax (1)
Train	2319	741
Validation	774	247
Test	774	247

Table 1: Class-wise distribution across dataset splits.

The distribution has been demonstrated in Table 1. To handling the class imbalance problem, we used different preprocessing techniques. The details procedures have been demonstrated in the Methodology section.

4 Methodology

This section provides an overview of the methodology and approach that have been used to build the system using the previously described dataset and transformer model. The methodology of our work

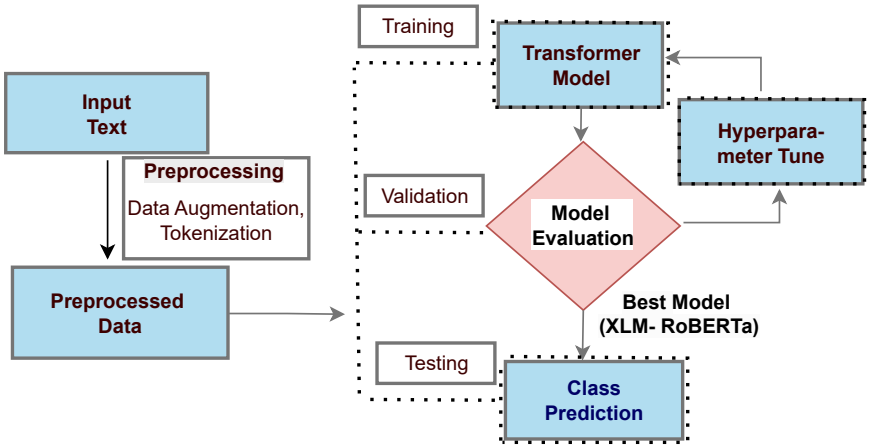

Figure 2: Methodology of our work

is shown in Figure 2.

4.1 Preprocessing

In our approach, preprocessing has focused on handling the unique challenges of code-mixed Hindi-English. The dataset is evaluated to determine its distribution and structure. The labels are encoded as containing Racial Hoax: 1, Not Racial Hoax: 0. The provided dataset was label encoded. We began by the tokenization and encoding the text using the AutoTokenizer from the HuggingFace library. For this task we used hing-roberta for Hindi-English code-mixed language. Text augmentation was applied during training, with a 10% random word masking. Additionally, the dataset was split into training and validation sets with stratified sampling to ensure balanced class distribution. The dataset is divided into training and validation sets using an 80%- 20% ratio.

4.2 Model Architecture and Training

For model selection, we have chosen XLM-RoBERTa, a multilingual transformer-based model, to capture the diverse nature of Hindi-English code-mixed data. We utilized XLM-R as the base model due to its proven effectiveness in multilingual tasks and its strong performance in low-resource languages. Prior work in deception detection, like Biradar et al. (2024a), has used similar transformer-based models successfully. Moreover, XLM-R's ability to capture cross-lingual semantic nuances makes it well-suited for hoax speech detection, which often relies on code-mixing.

Then, the provided dataset is converted into the Hugging Face Dataset format. The model architecture was improved with a custom classifier head that features dropout, layer normalization, and ReLU activation. Initially, all layers of the model were frozen. During the training period, we applied gradual unfreezing, starting with the last two layers. We optimized the model using AdamW with differential learning rates. Early stopping is implemented to prevent overfitting by monitoring validation loss.

4.3 Evaluation and Testing

During model evaluation, we assessed performance using the new dataset for development to fine-tune hyperparameters and ensure optimal performance. Once the model had achieved satisfactory results, we proceeded with the test dataset for the final classification. We have utilized the test dataset that has been provided by the shared task competition, which contains unlabeled comments, to classify racial hoax and non-racial hoax comments. The trained model predicts the labels, distinguishing between racial hoax and non-racial hoax content. This ensured the model's ability to generalize effectively to unseen data.

5 Result and Error Analysis

In this section, we have compared the results and analyzed the different transformer's performance based on the evaluation metrics. The macro F1-score measures the supremacy of the models. Table-3 shows the evaluation metrics for our best model.

5.1 Parameter Setting

We have tuned different hyperparameters to find the corresponding transformer's best model. The

Hyperparameter	Value
Learning Rate	5e-4 (Max)
Batch Size	16
Epochs	10
Dropout	0.3
Weight Decay	0.01
Masking Prob.	0.1
Optimizer	Adam
Scheduler	OneCycleLR
Early Stopping	Patience= 03 epochs

Table 2: Hyperparameters of the model

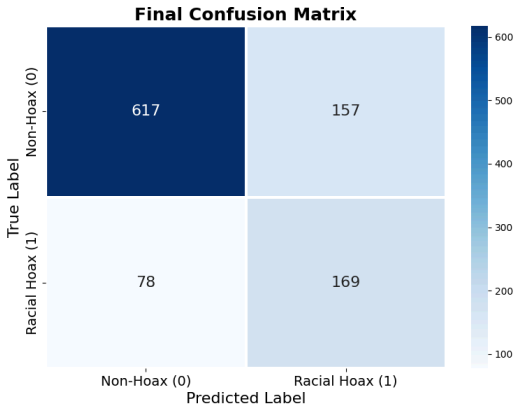

Figure 3: Confusion matrix of our transformer model

hyperparameters that are used in our model are shown in Table 2.

5.2 Metrics Evaluation

The performance of different models are evaluated by various metrics such as F1-score, Accuracy, Precision, and Recall on the test (Provided dev dataset) set. This ensured the model's ability to generalize effectively to unseen data. The evaluation metrics of our best model (XLM- RoBERTa) are shown in Table- 3. Figure- 3 demonstrates the confusion matrics of our model. Both of Table-3 and Figure 3 are based on the validation dataset. So we have an f1 score of 0.81 for the validation dataset, where the test data (unseen data) has an f1 score of 0.75.

5.3 Error Analysis

An analysis of the dataset splits has revealed a consistent class imbalance across the training, validation, and test sets. In each split, approximately 75.8% of the samples belong to the Non-Hoax class, while only 24.2% correspond to the Racial Hoax

Evaluations	Value
F1-Score	0.81
Val. Loss	0.183
Accuracy	85.57%
Precision	0.87
Recall	0.75

Table 3: Evaluation Metrics on the validation set

class. That's why the result is skewed to the non-racial hoax due to class imbalance problem. After analyzing the error, we found that the racial hoax containing text is misclassified as a non-racial hoax in some cases.

The confusion metrics show them well. These are due to the language morphology and lexical ambiguity, sarcasm, and irony when the context of the sentence is ambiguous. Rare words or dialects may also be another reason for these misclassifications. The evaluation metrics of our best model for corresponding languages are shown in Table 3. Incorporating additional context using hierarchical models could help in better understanding the context. Fine-tuning multilingual transformers in domain-specific corpora may also improve performance.

6 Conclusion

In this study, we proposed a transformer-based classification pipeline for detecting racial hoaxes in code-mixed Hindi-English social media content. Our approach incorporated robust preprocessing, Hinglish-specific tokenization, and fine-tuned multilingual models such as XLM-RoBERTa. Due to the problem of class imbalance in the dataset, we placed additional emphasis on preprocessing, incorporating techniques such as oversampling, data augmentation to mitigate the skew and enhance model generalization. That is why the experimental results demonstrated the effectiveness of our methodology, achieving strong performance on the LT-EDI@LDK-2025 shared task. Among all international participants, our system secured first place in the shared task with a satisfactory F1-score of 0.75, demonstrating the effectiveness of our method in addressing racially manipulative content in multilingual contexts. These findings highlight the importance of addressing racial disinformation in multilingual online spaces using deep learning.

Limitations

While our approach demonstrates better performance, it has certain limitations also. First of all, the provided dataset is quite small and class imbalance problem. The impact of the dataset on model development is visible in the result and error analysis section. The class imbalance problem skews the expected output to non-racial hoax class. Improving the dataset volume and more sample for 'Racial Hoax' class, better output can be expected. Secondly, our model shows limitations in capturing the sarcasm, irony, or implicit abusive content. As these are low resources languages and due to their native morphology, capturing the context is challenging.

References

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE symposium on security and privacy*, pages 461–475. IEEE.

Zo Ahmed, Bertie Vidgen, and Scott A Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, 11(1):8.

Luis Joaquín Arellano, Hugo Jair Escalante, Luis Villaseñor Pineda, Manuel Montes y Gómez, and Fernando Sanchez-Vega. 2022. Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish.

Kim Barker and Olga Jurasz. 2021. Text-based (sexual) abuse and online violence against women: Toward law reform? In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 247–264. Emerald Publishing Limited.

Shankar Biradar, Kasu Sai Kartheek Reddy, Sunil Saumya, and Md. Shad Akhtar. 2024a. Proceedings of the 21st international conference on natural language processing (ICON): Shared task on decoding fake narratives in spreading hateful stories (faux-hate). In *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*, pages 1–5, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2024b. Faux hate: unravelling the web of fake narratives in spreading hateful stories: a multi-label and multi-class dataset in cross-lingual hindi-english code-mixed text. *Language Resources and Evaluation*, pages 1–32.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 386–393.

Concetta Papapicco, Isabella Lamanna, and Francesca D'Errico. 2022. Adolescents' vulnerability to fake news and to racial hoaxes: A qualitative analysis on italian sample. *Multimodal Technologies and Interaction*, 6(3):20.

Md Mizanur Rahman, Srijita Dhar, Md Mehedi Hasan, and Hasan Murad. 2025. MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL based transformer models for detection of abusive Tamil and Malayalam text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 243–247, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

CUET's_White_Walkers@LT-EDI-2025: A Multimodal Framework for the Detection of Misogynistic Memes in Chinese Online Content

Md Mubasshir Naib[a], Md Mizanur Rahman[b], Jidan Al Abrar[c]
Md Mehedi Hasan[d], Md Siddikul Imam Kawser [e], Mohammad Shamsul Arefin [f]
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904089[a], u1904116[b], u1904080[c],u1904067[d],u1904081[e]}@student.cuet.ac.bd,
sarefin@cuet.ac.bd[f]

Abstract

Memes, combining visual and textual elements, have emerged as a prominent medium for both expression and the spread of harmful ideologies, including misogyny. To address this issue in Chinese online content, we present a multimodal framework for misogyny meme detection as part of the LT-EDI@LDK 2025 Shared Task. Our study investigates a range of machine learning (ML) methods such as Logistic Regression, Support Vector Ma- chines, and Random Forests, as well as deep learning (DL) architectures including CNNs and hybrid models like BiLSTM-CNN and CNN-GRU for extracting textual features. On the trans- former side, we explored multiple pretrained models including mBERT, MuRIL, and BERT- base-chinese to capture nuanced language rep- resentations. These textual models were fused with visual features extracted from pretrained ResNet50 and DenseNet121 architectures using both early and decision-level fusion strategies. Among all evaluated configurations, the BERT-base-chinese + ResNet50 early fusion model achieved the best overall performance, with a macro F1-score of 0.8541, ranking 4th in the shared task. These findings underscore the effectiveness of combining pretrained vision and language models for tackling multimodal hate speech detection.

1 Introduction

In recent years, meme culture has become a dominant form of expression on Chinese social media platforms. Memes are often humorous or satirical, but like any medium, they are not immune to misuse (Das et al., 2022). Increasingly, this format is being exploited to disseminate discriminatory or hateful ideas—misogyny among the most concerning. Identifying such harmful content is essential in creating safer digital communities.

While the detection of offensive content in memes has seen growing attention (Mohiuddin et al., 2025), research that targets misogynistic content specifically, particularly in the Chinese context, is still in its infancy. Prior studies have largely focused on high-resource languages such as English, Hindi, and Arabic where comprehensive datasets and pretrained models are more readily available. By contrast, the Chinese language, despite its vast user base, remains comparatively underrepresented in this domain (Chowdhury et al., 2025).

To address this shortfall, we introduce a multimodal detection framework tailored to the nuances of Chinese meme content. Central to our approach is a newly curated dataset, annotated with misogynistic and non-misogynistic labels. The dataset reflects the linguistic and cultural diversity of Chinese online spaces, ensuring the model learns from contextually relevant examples.

Our methodology combines textual analysis, leveraging powerful transformer-based models to parse captions and embedded text. To complement textual analysis, we extract visual features using pretrained convolutional networks like ResNet50 and DenseNet121. These features are then integrated with text representations using early and decision-level fusion strategies. Our best-performing configuration, which combines BERT-base-chinese and ResNet50 via early fusion, demonstrates the effectiveness of this multimodal approach in capturing the complex and often subtle nature of misogynistic content in memes (Fersini et al., 2019).

Our main contributions are as follows:

- We develop and fine-tune multimodal models that integrate both textual and visual information.
- We evaluate multiple model configurations, offering insight into effective strategies for detecting nuanced hate speech in memes.

The implementation details have been pro-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 68–74

vided in the following GitHub repository:-https://github.com/MubasshirNaib/Misogyny-Meme-Detection.

2 Related Work

The rise of misogynistic(Hossan et al., 2025) and harmful content(Naib et al., 2025; Sakib et al., 2025) in online memes has sparked growing concern and has become a significant area of research. As these memes typically combine both text and images, researchers have increasingly turned to multimodal learning techniques to improve detection capabilities. These techniques aim to process and interpret both the visual and linguistic components of a meme simultaneously—a task that becomes particularly complex in the Chinese context, given its rich cultural references, nuanced language, and diverse writing systems.

Although studies directly targeting misogynistic memes in Chinese are limited, various recent works offer solid ground for building suitable approaches. For example, (Jindal et al., 2024) introduced the MISTRA model, which merges text features with image embeddings using variational autoencoders (VAEs) to compress the image data effectively. This fusion allows the system to capture deeper semantic correlations between text and visuals.

Expanding on this idea, (Srivastava, 2022) developed MOMENTA, a deep neural network that looks at both broad and detailed features within memes. By analyzing overall structure alongside fine-grained details, the model is better equipped to spot nuanced forms of hate speech or misogyny.

Addressing the multilingual nature of memes, (Singh et al., 2024) compiled a large code-mixed dataset in Hindi and English, aimed specifically at identifying misogynistic content. They showed that multimodal approaches, especially those trained on code-switched language, are better suited for the mixed-language realities of social media—an insight that is also applicable to Chinese content, which often blends Mandarin with dialects, slang, or romanized expressions.

A notable contribution by (Pramanick et al., 2021) is the SCARE framework, which emphasizes strong alignment between textual and visual data. The model works by maximizing mutual information across the two modalities, making their shared representation more cohesive and informative. At the same time, it refines how each modality is represented on its own.

Meanwhile, (Habash et al., 2022) took a different approach by combining multiple models into an ensemble. This method benefits from the strengths of each individual model, helping to offset their weaknesses and improve overall detection rates of misogynistic content.

The importance of linguistic and cultural diversity in meme detection was also highlighted in the DravidianLangTech-2022 shared task (Das et al., 2022), where teams focused on memes in languages like Tamil and Malayalam. Their findings reinforced the value of fusing image and text data, especially in low-resource languages. Supporting studies by (Ghanghor et al., 2021) and (Chakravarthi et al., 2024) echoed these results, offering evidence that multilingual, multitask frameworks can effectively capture offensive and misogynistic content across different languages and contexts.Despite these strides, Chinese memes remain an underexplored territory. The combination of symbolic imagery, character-based writing, sarcasm, and internet-specific language poses unique challenges. For any framework designed to detect misogyny in Chinese memes, it's crucial to handle visual-linguistic humor, character-level nuance, and even cultural cues that may not be obvious without context.

3 Task and Dataset Description

This study addresses the task of misogyny meme detection in Chinese social media as part of the LT-EDI@LDK 2025 Shared Task(Chakravarthi et al., 2025). The given dataset (Ponnusamy et al., 2024) is multimodal, comprising image-text meme pairs labeled as Misogyny or Not-Misogyny. It is divided into training (1190 samples), validation (170 samples), and a test set (340 samples). The training set includes 841 non-misogynistic and 349 misogynistic examples, while the validation set includes 123 and 47 respectively. The textual data consists of 4,553 total words and 3,902 unique words, reflecting rich linguistic diversity. This setup enables the development of multimodal models that capture both visual cues and nuanced language patterns essential for detecting gender-based harmful content. Table 1 shows the class-wise distribution of samples for the Chinese dataset.

Class	Train	Validation	Test	W_T	UW_T
Not-Misogyny	841	123	236	2216	1996
Misogyny	349	47	104	1373	1139
Total	1190	170	340	4553	3902

Table 1: Class distribution across training, validation, and test splits, where W_T represents total words and UW_T represents total unique words.

4 Methodology

Several ML, DL, and transformer-based models were investigated to construct a robust framework for misogyny meme detection (Figure 1).The implementation details of the models have been open-sourced to ensure reproducibility[1].Appendix A presents the system requirements.

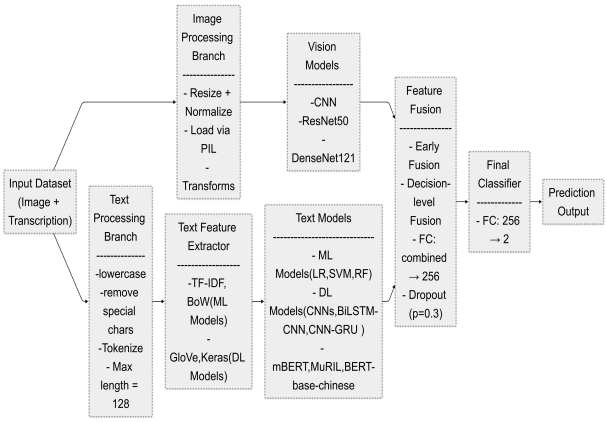

Figure 1: Schematic workflow for detecting misogynistic memes in Chinese social media content.

4.1 Data Preprocessing

The dataset comprises memes with both image and textual components. These are processed through two distinct branches to prepare them for feature extraction. In the image processing branch, all images are resized, normalized, and loaded using the Python Imaging Library (PIL). Further transformations, such as random cropping and flipping, are applied to enhance robustness and mitigate overfitting. On the textual side, transcriptions are first normalized by converting to lowercase and removing special characters. The cleaned text is then tokenized, with a maximum token length of 128, ensuring consistency in input dimensions for downstream models.

[1]https://colab.research.google.com/drive/1TYrg-vma1h46UwuhTdtX0eO2pCVPwJTn?usp=sharing

Hyperparameter	BERT-base-chinese + ResNet50 (Chinese)
Learning Rate	5e-5
Batch Size	8
Number of Epochs	5
Max Sequence Length	128
Optimizer	AdamW
Dropout Rate	0.3
Image Model	ResNet50
Text Model	BERT-base-chinese
Scheduler	ReduceLROnPlateau

Table 2: Tuned hyperparameters used in the best-performing multimodal model for Chinese misogyny meme detection.

4.2 Feature Extraction

Visual features are extracted using a range of pretrained convolutional neural networks, including generic CNNs, ResNet50, and DenseNet121. These models transform the image input into high-level visual embeddings. In parallel, textual features are extracted through two categories of approaches. Traditional machine learning pipelines utilize TF-IDF and Bag-of-Words representations. These features are suitable for models such as Logistic Regression, Support Vector Machines, and Random Forests. For deep learning models, word embeddings like GloVe and Keras embeddings are employed, supporting CNN-based architectures and hybrid models like BiLSTM-CNN and CNN-GRU. Additionally, transformer-based language models—mBERT, MuRIL, and BERT-base-chinese—are used to extract context-rich text representations. BERT-base-chinese, in particular, demonstrated superior performance in capturing linguistic nuances specific to Chinese discourse.

4.3 Baselines

To establish comparative performance, both unimodal and multimodal baselines were evaluated.

4.3.1 Unimodal Baselines

In the unimodal setup, the text and image modalities are processed independently. Text-based models include both machine learning classifiers using traditional features (TF-IDF, BoW) and deep learning architectures with embeddings. Similarly, image-based baselines rely on pretrained CNNs like ResNet50 and DenseNet121. Each modality is passed through its respective pipeline, and the final classification is performed separately to assess individual performance.

Approaches	Classifiers / Models	P	R	F1	G1
Textual Only	Logistic Regression	0.7721	0.6105	0.6453	0.6827
	SVM	0.7854	0.6282	0.6723	0.6985
	Random Forest	0.7668	0.5921	0.6302	0.6694
	CNN	0.7013	0.6450	0.6718	0.6729
	BiLSTM-CNN	0.7219	0.6523	0.6841	0.6855
	CNN-GRU	0.7352	0.6604	0.6907	0.6959
Visual Only	ResNet-50	0.7024	0.6201	0.6482	0.6595
	DenseNet-121	0.7480	0.6443	0.6827	0.6928
Multi-modal Fusion (Early Fusion)	mBERT + ResNet-50	0.8387	0.8034	0.8172	0.8209
	MuRIL + ResNet-50	0.8460	0.8127	0.8239	0.8292
	BERT-base-chinese + ResNet-50	**0.8812**	**0.8307**	**0.8541**	**0.8556**
Multi-modal Fusion (Late Fusion)	BERT-base-chinese + ResNet-50	0.8650	0.8201	0.8384	0.8421

Table 3: Comparison of various unimodal and multimodal models for misogyny meme detection in Chinese. EF: Early Fusion, LF: Late Fusion, P: Precision, R: Recall, F1: F1-score, G1: Geometric mean of P and R.

4.3.2 Multimodal Baselines

The multimodal approach fuses information from both text and image branches. Two primary fusion strategies are explored. Early fusion combines the intermediate feature representations from each modality and feeds them into a joint fully connected layer, followed by a dropout layer with a probability of 0.3 for regularization. In decision-level fusion, the output scores from unimodal classifiers are merged to generate the final prediction. Among all configurations, the early fusion model that integrates BERT-base-chinese and ResNet50 outperformed others, and the tuned hyperparameters are presented in Table 2.

5 Result Analysis

The performance evaluation of various unimodal and multimodal models on the Chinese misogyny meme detection task, which are presented in Table 3, reveals key insights into the effectiveness of different fusion strategies and model combinations. Among unimodal textual models, CNN-GRU and BiLSTM-CNN outperformed classical classifiers like Logistic Regression and SVM, demonstrating that sequential and convolutional architectures are better at capturing linguistic patterns. Visual-only models, particularly DenseNet-121, also performed reasonably well, though their standalone effectiveness remained slightly lower than that of text-based models.

Multimodal fusion approaches significantly outperformed unimodal methods. Early fusion models, especially BERT-base-chinese combined with ResNet-50, achieved the highest performance with an F1-score of 0.8541, indicating strong synergy between visual and textual features. Late fusion models also improved results but were slightly less effective than early fusion, emphasizing the value of integrating modalities early in the learning process. These findings affirm that combining vision and language models is crucial for accurately detecting misogynistic content in memes.Appendix B presents the error analysis.

6 Conclusion

In this study, we proposed a robust multimodal framework for misogyny meme detection in Chinese social media content, developed as part of the LT-EDI@LDK 2025 Shared Task. By combining pretrained transformer-based textual encoders such as BERT-base-chinese with visual feature extractors like ResNet-50 and DenseNet-121, our approach effectively captured the nuanced interplay between language and imagery. Among all tested configurations, the early fusion model of BERT-base-chinese and ResNet-50 achieved the best overall performance, demonstrating the strength of deep multimodal representation learning in tackling hate speech detection tasks. These results reinforce the importance of using culturally and linguistically aligned pretrained models for context-sensitive applications like misogyny detection.

7 Limitations

Despite promising results, our work has some limitations. First, the dataset was relatively small, particularly the test set, which may limit the generalizability of the findings. Second, the binary classification setting (misogyny vs. not-misogyny) does not capture the full spectrum or subtlety of harmful content. Third, although our fusion strategies improved performance, more sophisticated fusion

mechanisms such as attention-based or cross-modal transformers could further enhance the model's interpretability and accuracy. Finally, domain-specific biases in pretrained models and visual encoders may impact performance in culturally nuanced cases, calling for the development of more inclusive and fair AI systems.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025. An overview of the misogyny meme detection shared task for chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, Charmathi Rajkumar, and 1 others. 2024. Overview of shared task on multitask meme classification-unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144.

Md Sajid Alam Chowdhury, Mostak Mahmud Chowdhury, Anik Mahmud Shanto, Jidan Al Abrar, and Hasan Murad. 2025. Fired_from_nlp@ dravidianlangtech 2025: A multimodal approach for detecting misogynistic content in tamil and malayalam memes. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 459–464.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Hate-alert@ dravidianlangtech-acl2022: Ensembling multi-modalities for tamil trollmeme classification. *arXiv preprint arXiv:2204.12587*.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 222–229.

Mohammad Habash, Yahya Daqour, Malak Abdullah, and Mahmoud Al-Ayyoub. 2022. Ymai at semeval-2022 task 5: Detecting misogyny in memes using visualbert and mmbt multimodal pre-trained models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 780–784.

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET-NLP_Big_O@DravidianLangTech 2025: A multimodal fusion-based approach for identifying misogyny memes. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 427–434, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Md Mohiuddin, Md Minhazul Kabir, Kawsar Ahmed, and Mohammed Moshiul Hoque. 2025. Cuet-nlp_mp@ dravidianlangtech 2025: A transformer-based approach for bridging text and vision in misogyny meme detection in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 514–521.

Md. Mubasshir Naib, Md. Saikat Hossain Shohag, Alamgir Hossain, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. cuetRaptors@DravidianLangTech 2025: Transformer-based approaches for detecting abusive Tamil text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 739–745, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Nazmus Sakib, Md. Refaj Hossan, Alamgir Hossain, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET-NLP_Big_O@DravidianLangTech 2025: A BERT-based approach to detect fake news from Malayalam social media texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 440–447, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Harshvardhan Srivastava. 2022. Misogynistic meme detection using early fusion model with graph network. *arXiv preprint arXiv:2203.16781*.

A System Requirements

The entire framework is implemented using Python, leveraging libraries such as PyTorch and HuggingFace Transformers for deep learning, and Scikit-learn for traditional machine learning models. A GPU-enabled system with at least 16 GB of RAM is recommended to efficiently train and evaluate deep models, especially those involving transformer architectures and multimodal fusion.

B Error Analysis

We conducted both quantitative and qualitative error analyses to gain comprehensive insights into the performance of the proposed model.

B.1 Quantitative Analysis:

The confusion matrix B.1 reveals that while the model effectively identifies non-misogynistic content with a high specificity of 96.74% (119 true negatives vs. 4 false positives), it struggles to detect misogynistic content, as evidenced by a lower sensitivity of 59.57% (28 true positives vs. 19 false negatives). This indicates that the model is prone to under-detecting misogynistic content, potentially due to subtle or implicit cues that are not effectively captured by the current multimodal fusion strategy. Misclassification of misogynistic content as non-misogynistic suggests that the textual and visual features may not be sufficiently aligned, particularly in cases where misogyny is conveyed indirectly or through ambiguous visual elements. Addressing these gaps could involve refining feature extraction, implementing more targeted attention mechanisms, and expanding the training set with diverse and nuanced misogynistic examples.

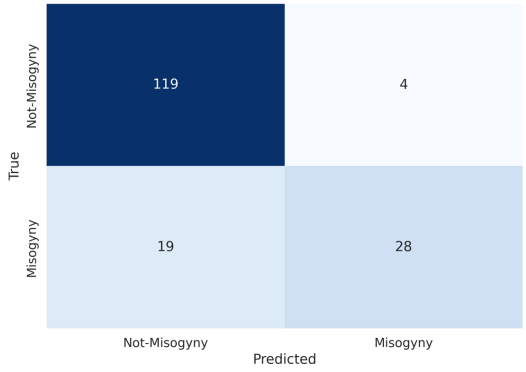

Figure B.1: Confusion matrix of the proposed model

B.2 Qualitative Analysis:

The qualitative analysis B.2 shows that the model correctly identifies non-misogynistic content in Sample 1 and Sample 4, proving that it can recognize harmless content even when the visuals seem aggressive, like the fire scene. In Sample 2, the model correctly detects obvious misogynistic content expressed through clear text, showing that it can identify direct misogynistic messages. However, in Sample 3, the model wrongly labels misogynistic content as non-misogynistic, suggesting that it struggles to detect more subtle or hidden forms of misogyny, especially when sarcasm or cultural references are used. To reduce such errors, the model could be improved by training it with more examples of subtle and indirect misogynistic content and by using attention mechanisms to focus on specific regions or words that convey implicit biases. Strengthening the alignment between textual and visual features could also help in capturing nuanced cues more effectively.

Image Id: 1582.jpg

True Label: Not-Misogyny(0)

Predicted Label: **Not-Misogyny(0)**

Sample 1

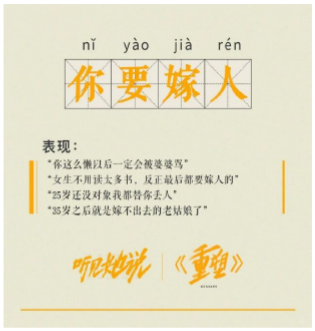

Image Id :1342.jpg

True Label: Misogyny(1)

Predicted Label: **Misogyny(1)**

Sample 2

Image Id: 933.jpg

True Label: Misogyny(1)

Predicted Label: **Not-Misogyny(0)**

Sample 3

Image Id :1305.jpg

True Label: Not-Misogyny(0)

Predicted Label: **Not-Misogyny(0)**

Sample 4

Figure B.2: Some outputs predicted by the best model.

CUET's_White_Walkers@LT-EDI 2025: Transformer-Based Model for the Detection of Caste and Migration Hate Speech

Jidan Al Abrar, Md Mizanur Rahman, Ariful Islam,
Md Mehedi Hasan, Md Mubasshir Naib, Mohammad Shamsul Arefin
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{080, 116, 129, 067, 089}@student.cuet.ac.bd, sarefin@cuet.ac.bd

Abstract

Hate speech on social media is an evolving problem, particularly in low-resource languages like Tamil, where traditional hate speech detection approaches remain under developed. In this work, we provide a focused solution for cast and migration-based hate speech detection using Tamil-BERT, a Tamil-specialized pre-trained transformer model. One of the key challenges in hate speech detection is the severe class imbalance in the dataset, with hate speech being the minority class. We solve this using focal loss, a loss function that gives more importance to harder-to-classify examples, improving the performance of the model in detecting minority classes. We train our model on a publicly available labeled dataset of Tamil text as hate and non-hate speech. Under strict evaluation, our approach achieves impressive results, outperforming baseline models by a considerable margin. The model achieves an F1 score of 0.8634 and good precision, recall, and accuracy, making it a robust solution for hate speech detection in Tamil. The results show that fine-tuning transformer-based models like Tamil-BERT, coupled with techniques like focal loss, can substantially improve performance in hate speech detection for low-resource languages. This work is a contribution to this growing amount of research and provides insights on how to tackle class imbalance for NLP tasks.

1 Introduction

The sudden rise in social networking websites has come with the ever-mounting responsibility of monitoring and regulating harmful content, primarily hate speech. Hate speech is defined as any form of speech that incites violence or is directed against individuals on the basis of race, religion, gender, or any other quality and hence is an emerging menace to the cyber world. Though most of the hate speech detection research has focused on high-resource languages like English, hate speech detection in low-resource languages is not yet explored. Tamil, a Dravidian language with millions of speakers, is a typical example of a low-resource language where hate speech detection tools are nonexistent or inefficient. Recognizing hate speech in Tamil is particularly challenging due to its complex syntax, local dialects, and lack of massive annotated datasets.

Latest advances in natural language processing (NLP) have revealed that transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) represent an effective remedy for text classification issues, such as hate speech recognition. But one of the persistent problems in training such models for hate speech detection is class imbalance. In the majority of datasets, hate speech instances are significantly less than non-hate speech instances, which may lead to model bias and bad generalization.

To address these issues, we propose a novel Tamil hate speech detection method based on Tamil-BERT, the language-specific variant of BERT. We enhance the model's performance with the help of focal loss, which is a technique that allows the model to focus more on the minority class and therefore mitigate the effects of class imbalance. Our contributions are as follows:

1. Developing a deep learning framework for Tamil hate speech detection using a Pre-trained transformer model.

2. Addressing class imbalance with focal loss, improving the detection of minority hate speech instances.

3. Evaluating the model on the available Tamil hate speech dataset with a competitive performance of F1 score 0.8634.

The implementation details have been provided in the following GitHub repository:-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 75–79

https://github.com/Mizan116/LT-EDI-LDK-2025/Hate Speech.

This study builds on the earlier studies of (Chhabra, 2022), who employed transformer models for hate speech detection across multiple languages, and (Zhao, 2020), who demonstrated the efficacy of BERT across low-resource languages. Our study goes one step further by applying these approaches to the special case of Tamil, demonstrating how fine-tuning transformer-based models and focal loss can be used to overcome the specific challenges posed by language-specific characteristics and dataset skew.

2 Related Work

Hate speech detection has attracted significant interest as a consequence of the rapidly rising amount of toxic material on social media platforms. Early approaches to coping with this problem used typical machine learning methods like Support Vector Machines (SVMs) and Naive Bayes (Waseem and Hovy, 2016), where manually designed features like n-grams were common. However, these approaches often struggled to accommodate the complex linguistic and contextual nature of hate speech.

With the arrival of deep learning, models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) emerged, with better performance in classification tasks, especially in recognizing sequential dependencies (Zhang et al., 2018). Yet these models were not yet able to deal with long-range dependencies and deeper context in text, especially in very morphologically complex languages such as Tamil.

The recent progress in Natural Language Processing (NLP) has been transformer-based models, such as BERT (Devlin et al., 2019), which has been the key driver for text classification. Models based on BERT have particularly excelled at hate speech detection since they can learn contextual information through attention mechanisms (Zhang et al., 2020). Other than this, experimentation has also demonstrated that multilingual BERT models can be used for hate speech detection in low-resource languages such as Hindi and Tamil (Chhabra, 2022), thereby demonstrating that transformer models can also be fine-tuned for low-resource languages.

Class imbalance is the greatest challenge in hate speech classification. It has been addressed by focal loss (Lin et al., 2017) that focuses on hard-to-class samples more, thus improving the identification of minority classes like hate speech. Focal loss has shown great promise for its application in NLP such as sentiment analysis and hate speech classification.

In addition, research involving adversarial training (Ta, 2022) and data augmentation strategies like paraphrasing (Bora, 2022) has indicated improvements in model robustness and performance in detecting aggressive language on social media.

This research is grounded on these breakthroughs and uses a Tamil-specific BERT model (Tamil-BERT) and focal loss to address class imbalance in hate speech detection in Tamil. In a related shared task on Dravidian languages, the authors of Rahman et al. (2025) employed transformer models like XLM-R and MuRIL, demonstrating high performance in abusive language detection.

3 Dataset

We have applied the given dataset for Shared Task on Caste and Migration Hate Speech Detection in LT-EDI@LDK 2025 (Rajiakodi et al., 2025), which is focused entirely on caste and migration hate speech detection in Tamil language (Ponnusamy et al., 2024). The dataset is of type two classes: Caste/Migration-related Hate Speech and Non-Caste/Migration-related Hate Speech. The dataset is divided into training, development, and test datasets. The training dataset contains 2,790 samples, and the validation and test datasets contain 598 samples each.

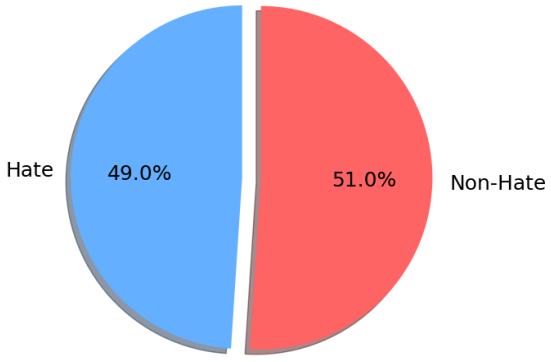

Figure 1: Class Distribution of Train Data

We used the dataset of Rajiakodi et al. (2024) for training, testing and evaluation our model. The distribution of the dataset is as follows:

Split	Hate	Non-Hate	Total Samples
Train	1,366	1,424	2,790
Dev	278	320	598

Table 1: Training and Development Split

Split	Hate	Non-Hate	Total Samples
Test	278	320	598

Table 2: Test Split

The information was collected from Tamil social media, with real-world forms of caste and migration-based hate speech. As there exists class imbalance, whereby non-hate speech instances are greater than hate speech samples, we used focal loss when training the model to give more importance to the minority class.

Text was tokenized using a Tamil-specific tokenizer and padded to 128 tokens. The evaluation metric for the task is macro F1-score to ensure balanced evaluation of both classes.

4 Methodology

In this section, we provide an overview of the methodology and approaches utilized to build the system using the previous Tamil-BERT transformer model. Methodology of our work is shown in Figure 2.

4.1 Preprocessing

The dataset used in this study consists of Tamil language social media texts annotated with binary labels as Hate: 1 and Non-Hate: 0. Preprocessing is crucial to ensure that the model receives clean and consistent input.We have done text cleaning, label encoding, data splitting and tokenization. In data splitting the data split into 80% training and 20% validation. Padding and truncating sequences to a maximum length of 128 tokens.

4.2 Model Selection

We selected Tamil-BERT for hate speech classification due to its strong language specific capabilities and superior performance in Tamil language.The Tamil-BERT model have the highest accuracy, precision, recall and F1-score compared to the other baseline models. The model was fine-tuned using cross-entropy loss and Adam optimizer to avoid overfitting.

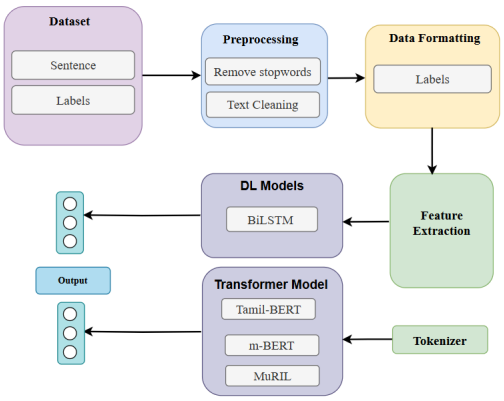

Figure 2: Methodology of our work

4.3 Evaluation and Testing

The Tamil-BERT model was evaluated on a 20% test set using Accuracy, Precision, Recall and F1-Score.A confusion matrix was used for error analysis.Tamil-BERT outperformed baseline models, showing strong generalization and reliable performance.Metrics were computed using macro-averaging to handle class imbalance, ensuring fair evaluation across categories.

5 Results and Analysis

In this section, we evaluate the Tamil-BERT model for hate speech binary classification through four components : parameter setting, comparative analysis, performance metrics and error analysis.

5.1 Parameter Setting

Table 3 shows parameter setting for Tamil-BERT model. In Table 3, lr, optim, bs and wd represent `learning_rate`, optimizer, `batch_size` and `weight_decay` respectively.

Model	lr	optim	bs	ep	wd
Tamil-BERT	2e-5	AdamW	16	8	0.05
MuRIL	3e-5	AdamW	32	7	0.1
m-BERT	2e-5	AdamW	32	5	-
BiLSTM	2e-5	Adam	16	8	-

Table 3: Parameter Setting in different model

5.2 Comparative Analysis

To validate the effectiveness of the proposed Tamil-BERT model, we compared its performance with several baseline models such as MuRIL, m-BERT and BiLSTM. To ensure robustness, we trained all

models across five different random seeds. The mean F1-score for Tamil-BERT was 0.8634, which consistently outperformed all baselines. Each model was trained on the same dataset and evaluated under identical conditions using macro-averaged metrics: Accuracy, Precision, Recall and F1-score.

The result, summarized in Table 4, shows that Tamil-BERT significantly outperforms the other models across all metrics. Here Loss, A, F1 denotes Loss, Accuracy and F1-Score. This is expected as Tamil-BERT is pre-trained specifically for the Tamil language and caste-related content.

Model	Loss	A	F1
Tamil-BERT	0.3571	87.22%	0.8634
MuRIL	0.4725	81.4%	0.77
m-BERT	0.5093	79.8%	0.74
BiLSTM	0.5761	74.2%	0.69

Table 4: Comparison of different models

5.3 Performance Metrics

The performance of various models has been evaluated using various metrics such as Accuracy, F1 Score, Precision, Recall and Confusion Matrix. Figure 3 show the confusion matrics of Tamil-BERT model for Tamil language.

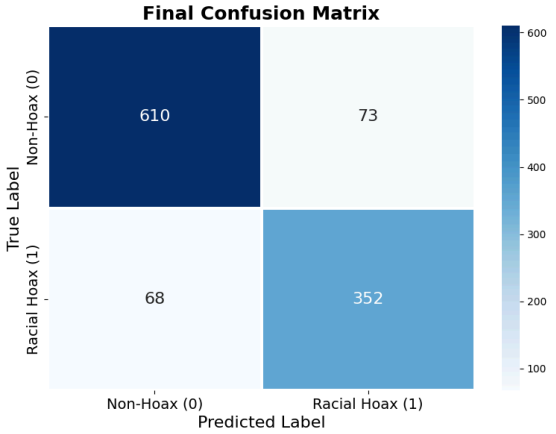

Figure 3: Confusion matrix of Tamil-BERT model

5.4 Error Analysis

The confusion matrix of Tamil-BERT model shows that the model correctly classified 610 negative and 352 positive instances with 73 false positives and 68 false negatives. The overall accuracy of Tamil-BERT is approximately 87.31% and the model achieves precision of 86.99%, recall of 85.85% and an F1-score of 86.34% for the positive class. The other baseline model have lower accuracy, lower precision and recall compare to the Tamil-BERT model. This error occur due to ambiguous or sarcastic language, code-mixed and formal text, implicit hate speech that is expressed indirectly.We also performed an ablation analysis by removing focal loss from the traingin pipleline and replace it with cross-entropy loss. This led to drop of 4.3% in macro F1-score, confirming that focal loss contributes significantly to model performance.

6 Conclusion

In this paper, we proposed a deep learning-based approach for caste and migration hate speech detection in Tamil based on the pre-trained transformer model Tamil-BERT. Using focal loss to handle class imbalance, our method attained a high F1-score of 0.8634, which testifies to its efficiency in classifying hate speech and non-hate speech. Although the model achieved promising performance, there is still room for further improvement, specifically in reducing overfitting and enhancing generalization. Future work will explore added features, better loss functions, and hyperparameter tuning to continue to advance performance. Our study contributes another entry to the growing corpus of work in low-resource hate speech detection and calls for responsible AI innovation to create secure digital spaces.

References

A. Bora. 2022. Data augmentation for hate speech detection using paraphrasing and back-translation. *Proceedings of the 2022 Annual Conference on Natural Language Processing*.

A. Chhabra. 2022. Hate speech detection using transformers: A comparative study. *Conference Name*.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*.

T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection. *Proceedings of ICCV 2017*.

Rahul Ponnusamy, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sathiyaraj Thangasamy, and Charmathi Rajkumar. 2024. Overview of Shared Task on

Caste/Immigration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Md Mizanur Rahman, Srijita Dhar, Md Mehedi Hasan, and Hasan Murad. 2025. MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL based transformer models for detection of abusive Tamil and Malayalam text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 243–247, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

H. T. Ta. 2022. Gan-bert: Adversarial learning for aggressive text detection on social media. *Proceedings of the 2022 IEEE/ACM International Conference on Computer-Aided Design*.

Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of NAACL-HLT 2016*.

Q. Zhang, K. Zhao, and X. Li. 2020. Bert for hate speech detection: A comparative study. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Z. Zhang, Y. Zhao, and Y. LeCun. 2018. Deep learning for hate speech detection. *Proceedings of the International Conference on Learning Representations (ICLR)*.

D. Zhao. 2020. Multilingual bert for hate speech detection. *Journal Name*.

NS@LT-EDI-2025: Caste/Migration based hate speech Detection

Nishanth S, Shruthi Rengarajan, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4aie22149, cb.en.u4aie22154}@cb.students.amrita.edu
s_sachinkumar@cb.amrita.edu

Abstract

Hate speech directed at caste and migrant communities is a widespread problem on social media, frequently taking the form of insults specific to a given region, coded language, and disparaging slurs. This type of abuse seriously jeopardizes both individual well-being and social harmony in addition to perpetuating discrimination. In order to promote safer and more inclusive digital environments, it is imperative that this challenge be addressed. However, linguistic subtleties, code-mixing, and the lack of extensive annotated datasets make it difficult to detect such hate speech in Indian languages like Tamil. We suggest a supervised machine learning system that uses FastText embeddings specifically designed for Tamil-language content and Whisper-based speech recognition to address these issues. This strategy aims to precisely identify hate speech connected to caste and migration, supporting the larger endeavor to reduce online abuse in low resource languages like Tamil.

Keywords: Hate Speech, Machine Learning models, FastText

1 Introduction

The rapid and sudden expansion of social media websites has transformed and revolutionized how people communicate with each other, exchange different types of information, and interact with a vast spectrum of disparate communities and communities of interest across the world (Sharma et al., 2025). Nevertheless, it also needs to be observed that such cyberspaces have also turned into important hotspots and breeding grounds for spreading hate speech, with the focus being specifically placed on targeting vulnerable communities, i.e., marginalized caste groups and migrant communities (Barman and Das, 2023)(Jahan and Oussalah, 2023). This malicious kind of abuse usually tends to manifest itself in the form of region-based abuses, coded language imparting veiled messages, and a range of derogatory statements that are employed to maintain prevailing social hierarchies and legitimize systemic discrimination (Kumar et al., 2017). The negative effects and negative implications stemming from such toxic content extend far and beyond the specific harm that results, and pose a very realistic threat to social cohesion, community harmony, and overall inclusivity in society(V P et al., 2023).

Although a great deal of research has been conducted and a great deal of investments have been made towards hate speech detection in high-resource languages like English, much remains to be done in addressing this very serious problem in low-resource languages properly (Papcunová et al., 2023)(MacAvaney et al., 2019) (K et al., 2021). This is a very acute problem in linguistically diverse and rich countries like India, where there are numerous languages and dialects spoken (Chakravarthi et al., 2023). Tamil, being one of the major languages of South India, has very distinctive problems concerning this, mainly because of its very unique linguistic properties, difference among various dialects, and the widespread use of code-mixing by the speakers. These cumulative problems make it particularly challenging to create useful automated hate speech detection systems. Moreover, the unavailability of properly annotated datasets only serves to further add to the issues of creating useful detection systems for this kind of abusive language (Rajiakodi et al., 2025).

In an effort to efficiently solve and address these major challenges, the Shared Task on Caste and Migration Hate Speech Detection, to be hosted at the well-regarded LT-EDI@LDK 2025 conference, has been crafted with the specific goal of encouraging and enabling large-scale research and development of robust machine learning models for addressing this major and urgent issue. The new proposed solution exploits FastText-based embeddings that

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 80–83

have been specifically designed for Tamil text processing and analysis. With a robust emphasis on hate speech detection and solving based on caste and migration challenges, this major task not only progresses but also fits within the wider and more general goal of creating safer, more respectful, and more inclusive digital spaces. This is especially critical for communities that use low-resource languages, which are likely to be exposed to special challenges and vulnerabilities in the digital space.

More details about the shared task can be found at[1].

2 Dataset

The dataset was distributed by the shared task organisers of Caste and Migration Hate Speech Detection - LT-EDI@LDK 2025 (Rajiakodi et al., 2024).

Data Type	Sample Size
Training	5512
Development (Dev)	787
Testing	1566

Table 1: Dataset split of the speech samples

The dataset used for this study comprises sentences in the Tamil language, categorized into two classes: Abusive and Non-Abusive (Class distribution). The data set is split into training and test sets. It was specifically developed for evaluating the suitability of language models for identifying abusive language in low-resource Dravidian languages, ensuring near-balanced Abusive and Non-Abusive example representations to provide more efficient training and evaluation.

3 Methodology

3.1 Data Preprocessing

The data pre-processing involves a series of steps such as conversion of data into lowercase, ensuring uniformity in the labels. The URLs and special characters are removed from the data to ensure consistency and to make sure the data is model friendly.

3.2 Embedding Generation

FastText offers character n-gram embeddings to enhance vector representation for morphologically rich languages. Words are represented as the average of these embeddings. It is a word2vec model extension. While FastText offers embeddings for character n-grams, the Word2Vec model offers embeddings for words. Similar to the word2vec model, fastText computes the vectors using CBOW and Skip-gram, using the subword information, allowing it to generate embeddings for out-of-vocablury words. This is especially useful for morphologically rich languages like Tamil.

In the proposed methodology, a 300-dimensional word embedding models for English and Tamil, trained on the Common Crawl Data is used.

The FastText model internally performs a series of functionalities. This includes:

- **Tokenization:** Each sentence is split into words.
- **Word Lookup:** For each word, the corresponding embedding vector is retrieved from the FastText model. Only vectors for words present in the model's vocabulary are kept.
- **Sentence Embedding:** The embedding of each sentence is computed by taking the mean of its word vectors, resulting in a fixed-size (300-dimensional) vector for each sentence.
- **Fallback for Empty Text:** If a sentence contains no valid words (e.g., only punctuation), a zero vector is assigned as its embedding.
- **Final Output:** A tensor of shape `[num_sentences, 300]` is returned, representing the sentence embeddings.

3.3 Machine Learning Models

The embeddings created are first loaded. Then, to make the final model more robust, the methodology combines the training and validation embeddings. This methodology performs feature-level fusion by concatinating the Tamil original embeddings and the Tamil-English translated embeddings so as to capture the semantics from both the languages in one input vector, enriching the feature space.

Several machine learning models have been used for this classification task (S et al., 2025):

- **XGBoost:** A powerful gradient-boosted tree model with GPU acceleration.
- **Logistic Regression:** A simple, linear classifier used as a baseline model.

[1] https://codalab.lisn.upsaclay.fr/competitions/21884

- **Random Forest:** An ensemble of decision trees that effectively handles feature interactions.

- **SVM (Support Vector Machine):** Particularly effective in high-dimensional spaces, making it suitable for text embeddings.

- **KNN (K-Nearest Neighbors):** A non-parametric method that bases predictions on the nearest neighbors.

- **MLPClassifier:** A shallow neural network (multi-layer perceptron).

4 Evaluation

The model is trained using the training embeddings and evaluated using the test embeddings. Metrics like the Accuracy, F1 score, Precision and Recall and the Training & Evaluation time has been taken into account. When all inferences of the models were submitted, the XGBoost model, which was trained in Tamil, emerged on top of others, securing us the rank **7th** with a F1 score of 0.80095

4.1 Final Model Performance

Dataset	Model	Acc	F1	Prec	Rec
Tamil	XGBoost	0.7878	0.7835	0.7859	0.7878
Tamil	Logistic Regression	0.6366	0.5697	0.6226	0.6366
Tamil	Random Forest	0.6607	0.5819	0.7026	0.6607
Tamil	SVM	0.6264	0.5088	0.6411	0.6264
Tamil	KNN	0.6163	0.6163	0.6163	0.6163
Tamil	MLP	0.7446	0.7457	0.7473	0.7446
Tamil-English	XGBoost	0.7700	0.7631	0.7685	0.7700
Tamil-English	Logistic Regression	0.6302	0.5675	0.6084	0.6302
Tamil-English	Random Forest	0.6595	0.5755	0.7120	0.6595
Tamil-English	SVM	0.6353	0.5369	0.6489	0.6353
Tamil-English	KNN	0.6455	0.6376	0.6357	0.6455
Tamil-English	MLP	0.7598	0.7604	0.7610	0.7598
Tamil (Orig+Eng)	XGBoost	0.7916	0.7860	0.7911	0.7916
Tamil (Orig+Eng)	Logistic Regression	0.6544	0.6150	0.6428	0.6544
Tamil (Orig+Eng)	Random Forest	0.6696	0.5882	0.7438	0.6696
Tamil (Orig+Eng)	SVM	0.6544	0.5941	0.6559	0.6544
Tamil (Orig+Eng)	KNN	0.6226	0.6154	0.6127	0.6226
Tamil (Orig+Eng)	MLP	0.7827	0.7829	0.7831	0.7827

Table 2: Performance of Various Models on Tamil and Tamil-English Datasets

The code files for this project can be accessed from[2]

[2] https://github.com/NishanthSaravanamurali/NS-LT-EDI-2025-Caste-Migration-based-hate-speech-Detection.git

5 Conclusion

This paper presents the results of a task performed as part of the Fifth Workshop on Speech and Language Technologies for Dravidian Languages, focusing on Caste and Migration Hate Speech Detectio in Tamil speech dataset. The conference provided the dataset for the proposed task. The proposed methodology makes use of the FastText model and embedding generation to train the models and compare the accuracies.

6 Limitations

While working on this topic, the major limitation we faced is the mixed language, as some text was in English and some was in Tamil. Translating one language causes loss of contextual information which in this case is important even though it provided additional features that can help ML classifiers help classify better.

References

Shubhankar Barman and Mithun Das. 2023. hate-alert@dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In *DRAVIDIANLANGTECH*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Sreelakshmi K, Premjith B, and Soman Kp. 2021. Amrita_CEN_NLP@DravidianLangTech-EACL2021: Deep learning-based offensive language identification in Malayalam, Tamil and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, Kyiv. Association for Computational Linguistics.

S. Sachin Kumar, M. Anand Kumar, and K. P. Soman. 2017. Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In *Mining Intelligence and Knowledge Exploration*, pages 320–334, Cham. Springer International Publishing.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019.

Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):e0221152.

Jana Papcunová, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex Intelligent Systems*, 9:2827–2842.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Nishanth S, Shruthi Rengarajan, S Ananthasivan, Burugu Rahul, and Sachin Kumar S. 2025. ANSR@DravidianLangTech 2025: Detection of abusive Tamil and Malayalam text targeting women on social media using RoBERTa and XGBoost. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 711–715, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar Singh. 2025. Hate speech detection research in south asian languages: A survey of tasks, datasets and methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(3).

Abeera V P, Dr. Sachin Kumar, and Dr. Soman K P. 2023. Social media data analysis for Malayalam YouTube comments: Sentiment analysis and emotion detection using ML and DL models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

SSN_IT_HATE@LT-EDI-2025: Caste and Migration Hate Speech Detection

Maria Nancy C[1], **Radha N** [2], **Swathika R**[3]
[1]Annai Veilankanni's College of Engineering, Nedungundram , India
[2,3] Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, India
nancycse13@gmail.com[1]
radhan@ssn.edu.in[2]
swathikar@ssn.edu.in[3]

Abstract

This paper proposes a transformer-based methodology for detecting hate speech in Tamil, developed as part of the shared task on Caste and Migration Hate Speech Detection. Leveraging the multilingual BERT (mBERT) model, we fine-tune it to classify Tamil social media content into caste/migration-related hate speech and nonhate speech categories. Our approach achieves a macro F1-score of 0.72462 in the development dataset, demonstrating the effectiveness of multilingual pretrained models in low-resource language settings. The code for this work is available on github Hate-Speech-Deduction.

1 Introduction

Hate speech poses a threat to marginalized communities, especially those affected by caste discrimination and migration. In India, these sensitive issues often fuel online hate, commonly expressed in regional languages like Tamil. Addressing such content is vital to fostering respectful digital spaces. Automated detection of hate speech in Tamil presents challenges due to its low-resource nature, complex morphology, frequent code-mixing with English, and informal writing style. Existing tools and datasets often prioritize high-resource languages, leaving Dravidian languages underrepresented. We propose a multilingual transformer-based system to identify caste- and migration-related hate speech in Tamil social media. Using a curated, annotated dataset, we fine-tune the bert-base-multilingual-cased model with BERT tokenization, cross-entropy loss, and evaluate performance via standard metrics. Predictions on unseen test data gauge generalization ability. This study contributes to ethical AI by addressing identity-based harm in underrepresented languages. By applying advanced NLP methods, we aim to promote safer, more inclusive online platforms for vulnerable groups.

2 Literature Survey

Hate speech detection has emerged as a vital area of natural language processing (NLP), focusing on identifying abusive, derogatory, or inciting content across multiple platforms and languages. Early work in this field primarily employed statistical machine learning models, such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression, with hand-crafted features like n-grams and TF-IDF vectors (Zampieri et al., 2020). With the advent of deep learning, models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) demonstrated superior performance, especially on noisy and informal texts common in social media (Badjatiya et al., 2017). However, these models often struggled with contextual understanding and multilingual settings. Transformer-based models like BERT (Devlin et al., 2019) revolutionized NLP by introducing contextualized embeddings and enabling transfer learning. These models improved the performance of hate speech classification in multiple languages. (Sutejo and Lestari, 2018) For low-resource languages, multilingual variants like mBERT and XLM-RoBERTa (Conneau et al., 2020) are particularly valuable. In the Indian context, (Bhattacharya et al., 2021) studied hate speech in Tamil and Malayalam, addressing challenges such as code-mixing, orthographic variation, and dialect diversity. (Patankar et al., 2022) developed transformer-based systems for abusive comment detection in Tamil, and (Roy et al., 2022) proposed a deep ensemble framework for detecting hate in multiple Dravidian languages. A landmark overview by (Rajiakodi et al., 2024) detailed the objectives, dataset structure, and methodological landscape of the LT-EDI shared task, with specific focus on caste and migration hate speech in Tamil. Comprehensive surveys, such as (Fortuna and Nunes, 2018), have discussed key limitations in early hate speech

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 84–89

research, advocating for more context-aware models. (Abro et al., 2020) emphasize the importance of robust datasets and suggest improvements in annotation quality and domain adaptation. Several recent works have emphasized ethical concerns, bias mitigation, and fairness in hate speech classifiers (Parker and Ruths, 2023). (Jahan and Oussalah, 2023) highlight ethical pitfalls and recommend model transparency and explainability. Multimodal approaches (Gomez et al., 2019);(Wu and Bhandary, 2020) combining text with images or audio have shown that non-textual features such as tone, pitch, and visual cues can enhance hate detection. While promising, these approaches are more resource-intensive and less feasible in text-only shared tasks. Further, enhancements in semantic understanding like sentiment integration (Zhou et al., 2021) and contextual embeddings (Malik et al., 2022) have contributed to improved classification accuracy. Techniques like SMOTE for balancing class distributions (Kovács et al., 2021) also play a critical role when dealing with imbalanced hate speech datasets. Taken together, these contributions form the basis of our methodological decisions for this shared task submission.

3 Proposed Methodology

3.1 Dataset Description

The dataset comprises Tamil-language texts collected from various social media platforms, reflecting real-world discourse and often containing informal, emotionally charged, or contextually nuanced language. Each entry in the dataset is annotated with a binary label indicating whether the content includes hate speech directed at caste or migration groups, or not. In total, the dataset includes 5512 training samples, 787 development samples, and 1576 test samples. These texts range from short phrases to longer posts, with many exhibiting informal spelling, colloquial expressions, and a high frequency of code-mixing between Tamil and English. Such linguistic diversity presents both opportunities and challenges for automatic classification. Our proposed system for caste and migration hate speech detection in Tamil is built upon a fine-tuned multilingual BERT (mBERT) model. We considered both mBERT and XLM-RoBERTa for this task, as both are widely used multilingual models that perform well on low-resource languages. However, we chose to work with mBERT because it is lighter and faster to train, which made it a better fit for our available resources. mBERT has also been shown to work well in similar tasks involving Tamil and other Dravidian languages. While XLM-RoBERTa might offer slightly better results in some cases, our initial experiments showed that mBERT still gave strong performance and was more efficient overall. Given these factors, we felt mBERT was the more practical choice for this study. The architecture includes multiple stages, starting from data preprocessing to model inference. This section elaborates on the overall workflow, including preprocessing, tokenization, model architecture, training, and evaluation.

3.2 Text Preprocessing

The first stage involves preprocessing the raw Tamil text from the dataset to standardize and clean the input. Since Tamil is a case-insensitive script, we did not apply any lowercasing, as it does not affect the language and may discard meaningful formatting in code-mixed text. Instead, we focused on removing URLs, mentions, hashtags, special characters, and redundant white spaces using regular expressions to clean the input without distorting its structure. However, we retained casing for English words in code-mixed content, as it may carry emphasis or mark named entities, which could be useful for classification. One of the key challenges in this task is the presence of code-mixed content, where users often switch between Tamil and English within a single sentence. In many cases, Tamil words are also transliterated using Roman script, making them harder to detect using standard tokenizers. In our current approach, we did not apply any special preprocessing for code-mixing or transliteration. Instead, we relied on the multilingual capabilities of mBERT, which is pretrained on multiple scripts and languages, including English and Tamil. While this provides some level of generalization, we acknowledge that the model may not fully capture the nuances of code-switched text or romanized Tamil. Additionally, redundant white spaces are stripped to produce cleaner and more consistent input for tokenization.

3.3 Tokenization

Once preprocessed, each text instance is tokenized using the bert-base-multilingual-cased tokenizer from the HuggingFace Transformers library. The tokenizer breaks the text into sub-word units, adds special tokens like [CLS] and [SEP], and generates input IDs, attention masks, and token type IDs. All

sequences are padded or truncated to a fixed maximum length of 256 tokens to ensure uniformity during batch processing.

3.4 Model Architecture

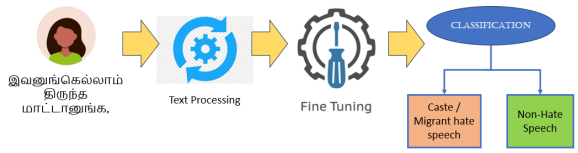

Figure 1: Workflow Diagram for Tamil Hate Speech Detection using mBERT

Figure 1 shows the proposed workflow of the model. We utilize the Bert for Sequence Classification model, which adds a classification head on top of the BERT encoder. The base model, bert-base-multilingual-cased, has 12 transformer layers and supports over 100 languages, including Tamil. The classification head is a fully connected layer that outputs two logits corresponding to the binary labels: caste/migration-related hate speech and nonhate speech.

3.5 Training Configuration

The model is fine-tuned in the labeled training set using a batch size of 32 and for 3 epochs. The optimizer used is Adam W with a learning rate of 2e-5 and weight decay to prevent overfitting.A linear learning rate scheduler was used to gradually reduce the learning rate during training. Although this scheduler supports warm-up, we did not apply it, as the number of warm-up steps was set to zero. The loss function used is CrossEntropyLoss, suitable for binary classification tasks. To facilitate efficient training and evaluation, the dataset is loaded using a custom PyTorch Dataset class and a Data-loader with shuffling enabled for the training set. Each batch is transferred to the GPU if available, ensuring accelerated computation.

3.6 Evaluation Strategy

After each epoch, the model is evaluated on the development set using a forward pass and argmax over output logits to generate predicted labels. Accuracy, precision, recall, and F1-score are calculated via the sklearn library to measure performance. In the final phase, the trained model predicts labels for the unseen test set, and results are saved in CSV format per the shared task protocol. This approach helps the model learn linguistic patterns and remain robust to the informal, noisy, and context-rich nature of Tamil social media. Leveraging mBERT's multilingual capabilities and fine-tuning on annotated domain-specific data, our system offers a practical solution for detecting hate speech in under-resourced languages. We evaluated model performance through experiments on the LT-EDI 2025 dataset, targeting hate speech against caste and migrant communities.

4 Experiment and Results

All experiments were carried out using PyTorch and Hugging Face Transformers on an NVIDIA Tesla V100 GPU, following the fine-tuning phase.

4.1 Evaluation Metrics

We used accuracy, precision, recall, and macro F1-score as our evaluation metrics. Among these, macro F1-score was prioritized due to the class imbalance and ethical weight of the task.

4.2 Development Set Results

Our best-performing model achieved a macro F1-score of 0.7246 on the development set. This indicated strong performance across both classes hate and non-hate speech despite the informal and code-mixed nature of the data.

4.3 Confusion Matrix

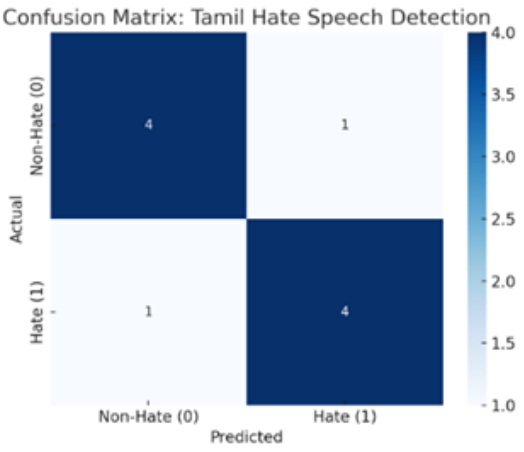

Figure 2: Confusion Matrix: Tamil Hate Speech detection against Caste / Migrated people

Figure 2 shows the model made balanced predictions across both classes. It successfully identified most hate speech posts while keeping false positives relatively low. The matrix also reveals some

instances where non-hate speech was incorrectly classified as hate, likely due to emotionally charged but non-derogatory language.

4.4 Class-wise Performance

In evaluating different machine learning approaches (refer to Table 1 and Table 2) for the detection of Tamil hate speech, both the Logistic Regression and Multinomial Naive Bayes models demonstrated moderate performance, with overall F1-scores of 0.66 and 0.64, respectively.

Class	Precision	Recall	F1-Score
Non-Hate (0)	0.68	0.70	0.69
Hate (1)	0.63	0.61	0.62
Accuracy			**0.66**
Macro Avg	0.66	0.65	0.65
Weighted Avg	0.66	0.66	0.66

Table 1: Logistic Regression Classification Report.

Class	Precision	Recall	F1-Score
Non-Hate (0)	0.71	0.67	0.69
Hate (1)	0.58	0.63	0.60
Accuracy			**0.65**
Macro Avg	0.65	0.65	0.64
Weighted Avg	0.66	0.65	0.65

Table 2: Multinomial Naive Bayes Classification Report.

Class	Precision	Recall	F1-Score
Non-Hate (0)	0.73	0.73	0.73
Hate (1)	0.72	0.72	0.72
Accuracy			**0.73**
Macro Avg	0.725	0.725	0.725
Weighted Avg	0.725	0.725	0.724

Table 3: BERT Model Classification Report.

The Logistic Regression model achieved slightly better balance between precision and recall across both classes, with a macro average F1-score of 0.65, while Multinomial Naive Bayes trailed close behind. These traditional models showed a tendency to perform better on the Non-Hate class, while struggling slightly with correctly identifying hate speech, as reflected in the lower F1-scores for class 1. In contrast, the transformer-based model (refer Table 3) significantly outperformed these baselines and achieved an F1-score of 0.7246. This improvement highlights the strength of deep learning architectures, especially in capturing complex linguistic patterns and contextual relationships that are common in nuanced languages like Tamil. The transformer model maintained balanced precision and recall across both classes, which contributed to its stronger overall performance compared to the other model. Based on this observation we can conclude that while traditional classifiers can serve as useful baselines, transformer-based models are far more effective for tasks requiring deeper semantic understanding, such as hate speech detection in low-resource languages.

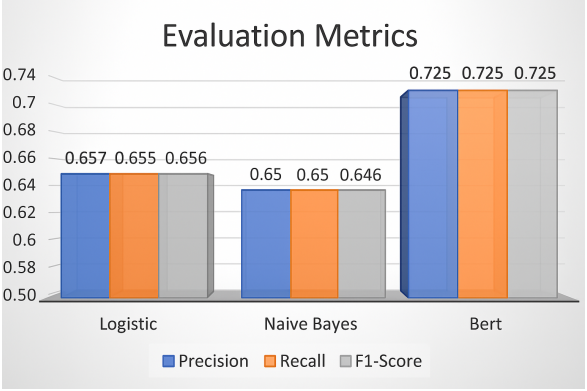

Figure 3: Evaluation Metrics using different methodologies

Figure 3 shows that among the models, BERT outperforms the others, achieving the highest values across all three metrics, with a consistent score of 0.725 for precision, recall and F1-score. Logistic regression follows, showing slightly lower but balanced scores: 0.657 for precision, 0.655 for recall, and 0.656 for F1-score. Naive Bayes, while close to Logistic Regression, performs slightly less effectively, particularly in F1-score, which stands at 0.646, compared to 0.65 for both precision and recall. This comparison highlights BERT as the most effective model in terms of classification performance for the given dataset.While the BERT-based model outperformed traditional classifiers with a macro F1-score of 0.7246, it's important to interpret these results in context. Detecting hate speech especially in low-resource, code-mixed languages like Tamil is inherently challenging due to informal language, slang, and subtle expressions of

bias. The score reflects moderate success, indicating that the model captures many hateful patterns but still struggles with nuanced or indirect speech. Therefore, this result should be viewed as a strong baseline rather than a final solution. Future work can build on this by incorporating linguistic context, domain-specific pretraining, or more advanced multilingual models.

4.5 Error Analysis

To get a better understanding of where the model struggles, we looked closely at some of the examples it got wrong in the development set. One of the most common issues was with posts that used sarcasm or indirect language to express hate. These types of messages didn't contain obvious offensive words, so the model often misclassified them as non-hate. Another challenge came from code-mixed posts especially those that switched between Tamil and English. In many cases, the hateful meaning was embedded in the Tamil part, but the English portion made the message sound neutral. This seemed to confuse the model. We also noticed that slang, spelling variations, and informal language, which are common on social media, made it harder for the model to correctly identify hate speech.In some cases, the model predicted hate where there was none. These false positives often included posts with strong emotions or criticism, but not targeted hate. The model likely relied on certain keywords or tone, misinterpreting emotional expression as harmful content.Overall, these errors show that while the model performs well on average, it still has trouble with nuance, subtlety, and cultural context especially in a language like Tamil. Understanding these mistakes not only helps explain the results but also points to areas we could improve in the future, like handling sarcasm, improving code-mixed understanding, or training with more context rich data.

5 Limitations

While our model performs well, it has notable limitations. Tamil social media posts often blend languages and include slang or informal expressions, which can confuse the model, especially when hate is subtly or sarcastically conveyed. The dataset used is small and only labels posts as hate or non-hate, overlooking the nuances in harmful expression. Since the model relies on mBERT, it struggles with cultural context and its predictions can be difficult to interpret. This raises concerns about fairness and bias, particularly if it learns problematic patterns from the training data. Social media often mirrors societal biases, which the model may unintentionally reinforce. Though we didn't perform an in-depth bias or fairness analysis in this study, exploring variations across caste, gender, or identity groups is a vital direction for future work. A deeper investigation into dataset and model biases could support fairer and more responsible deployment.

6 Conclusion

Our work shows that a multilingual model like mBERT can be fine-tuned to effectively detect caste and migration-related hate speech in Tamil social media posts. Even with limited data and the challenges of informal, mixed-language text, the model achieved good performance. This approach highlights the potential of using existing language models to support low-resource languages and address real social issues. We hope our method encourages more research in this area and helps make online spaces safer and more inclusive for everyone.

6.1 Future Work

In the future, we plan to explore multimodal approaches that combine text with audio or visual cues, as these could help capture more subtle or sarcastic forms of hate speech. We're also looking to expand our dataset and include more specific labels such as distinguishing between caste-based and migration-related content to enhance the model's accuracy. Additionally, we aim to incorporate cultural context, improve the explainability of model decisions, and address potential biases. These steps are crucial for building systems that are not only more accurate but also more trustworthy and practical for real world use.

References

S. Abro, S. Shaikh, Z. H. Khand, Z. Ali, S. Khan, and G. Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. In *Proceedings of the 2020 Conference on Hate Speech Detection*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Punyajoy Bhattacharya, Prateek Mishra, Indira Bhattacharya, and Amitava Das. 2021. Hate speech detection in low-resource languages: A case study on tamil and malayalam. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 93–101.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Ricardo Gomez, Michail Zervakis, and Björn Schuller. 2019. Multimodal hate speech detection: Combining text and visual features. In *Proceedings of the 2019 International Conference on Multimodal Interaction*, pages 602–606.

Md Shad Jahan and Mourad Oussalah. 2023. A review of nlp-based hate speech detection. *Information Processing & Management*, 60(2):102057.

Gábor Kovács, András Szabó, and Richárd Farkas. 2021. Addressing data scarcity in hate speech detection with external resources. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 25–34.

Junaid S Malik, N S Muralidhar, and Rakesh Tiwari. 2022. A comparative study of deep learning models for hate speech detection. *Procedia Computer Science*, 199:266–273.

Sage Parker and Derek Ruths. 2023. Assessing bias and fairness in hate speech detection systems. *ACM Transactions on the Web (TWEB)*, 17(1):1–23.

Siddhesh Patankar, Shubham Suryawanshi, and Bharathi Raja Chakravarthi. 2022. Abusive comment detection in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 66–72.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI 2024)*, Malta. EACL.

Prasenjit Kumar Roy, Soham Ghosh, and Animesh Mukherjee. 2022. Deep ensemble framework for hate speech detection in dravidian languages. In *Proceedings of the DravidianLangTech@ACL 2022*, pages 153–158.

Tony L Sutejo and Dwi P Lestari. 2018. Indonesian hate speech detection using deep learning. In *2018 International Conference on Asian Language Processing (IALP)*, pages 254–257. IEEE.

Chien-Sheng Wu and Ujjwal Bhandary. 2020. Hate speech detection in videos using multimodal cues. In *Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

Xinyu Zhou, Xilun Chen, and Yaqing Wang. 2021. Enhancing hate speech detection with sentiment knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 2764–2774.

ItsAllGoodMan@LT-EDI-2025: Fusing TF-IDF and MuRIL Embeddings for Detecting Caste and Migration Hate Speech

Amritha Nandini KL, Vishal S, Giri Prasath R, Anerud Thiyagarajan, Sachin Kumar S

Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{amrithanandini2003, vishalatmadurai, giriprasath017, anerud68511}@gmail.com, s_sachinkumar@cb.amrita.edu

Abstract

Caste and migration hate speech detection is a critical task in the context of increasingly multilingual and diverse online discourse. In this work, we address the problem of identifying hate speech targeting caste and migrant communities across a multilingual social media dataset containing Tamil, Tamil written in English script, and English. We explore and compare different feature representations, including TF-IDF vectors and embeddings from pretrained transformer-based models, to train various machine learning classifiers. Our experiments show that a Soft Voting Classifier that make use of both TF-IDF vectors and MuRIL embeddings performs best, achieving a macro F1 score of 0.802 on the test set. This approach was evaluated as part of the Shared Task on Caste and Migration Hate Speech Detection at LT-EDI@LDK 2025[1], where it ranked 6th overall.

1 Introduction

In India, caste and migration-based hate speech is a pervasive problem that has long been the focus of political discussion and still can be observed a lot in online forums and digital spaces. This carries severe real-world consequences, including psychological harm, the increase in social divisions, and the potential for inciting offline violence, particularly in linguistically diverse regions (Singh, 2025). This online discrimination takes place in numerous forms, including direct hate speech, derogatory remarks targeting specific castes and migration groups, cyberbullying of individuals based on their identity, threats of violence or social ostracism, and exclusion from online communities and groups.

With the boom of social media, such biases have found new avenues to spread, often under the disguise of free speech and anonymity. The spreading of hate speech targeting caste and migration groups not only normalizes discrimination but also reinforces harmful stereotypes, further marginalizing already vulnerable communities. Given the vast volume of online discourse and the rapid spread of harmful content, there is a pressing need to develop a system that is capable of recognizing and addressing such biases in real time.

Existing research on detection of hateful speech has primarily focused on widely spoken languages such as English. The complexity of code-mixed and regional language discourse prevalent in multilingual societies like India is often overlooked. Tamil, a widely spoken Dravidian language, frequently appears in code-mixed forms with English and other regional languages, making hate speech detection in Tamil code-mixed text a challenging task. The lack of sufficient annotated datasets with code-mixed text further complicate the problem.

Recent studies have explored a variety of transformer-based, machine learning, and data augmentation approaches for the detection of hate speech in Tamil, particularly in code-mixed and multilingual contexts. Using an ensemble of models such as XLM-RoBERTa, multilingual-cased BERT, and MuRIL, one of the best-performing models in the LT-EDI-EACL 2024 shared task achieved an F1-score of 0.82 (Singhal and Bedi, 2024). Another submission to this shared task evaluated 12 pre-trained transformer models in Indian multilingual language settings and found that MuRIL-Large was the most effective, with an F1-score of 0.81, which was obtained by ensembling the top-performing models (Pokrywka and Jassem, 2024). Another method experimented with transformers, FastText, and TF-IDF; mBERT did the best with an F1-score of 0.80 (Alam et al., 2024). To counter the difficulty of detecting masked abusive language in regional languages,(S et al., 2025) developed a system for Tamil and Malayalam by using supervised learning techniques on RoBERTa

[1] https://codalab.lisn.upsaclay.fr/competitions/21884

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 90–94

text embeddings. Researchers have also investigated multimodal hate speech detection that incorporates text, speech, and video data in addition to shared tasks. A study that compared several Tamil language models, such as Tamil-BERT, LaBSE, Hate-MuRIL, and MuRIL-Large-Cased, concluded that Tamil-BERT was the most successful (Mohan et al., 2025).

Another paper focused on detecting offensive language in Tamil-English code-switching by highlighting the potential of a hybrid system using both KANs and standard classifiers to improve detection accuracy (Jaidev et al., EasyChair, 2024). An averaging ensemble approach resulted in an accuracy score of 90.67% in identifying hate speech with mixed Tamil-English codes by using conventional machine learning approaches such as Support Vector Machine, Naive Bayes and ensemble methods (FHA et al., 2023). Different neural networks were explored, out of which a hybrid CNN-BiLSTM model adjusted for data imbalance, performed best for identifying offensive language in Dravidian languages (K et al., 2021).

Despite the substantial social impact of hate speech related to caste and migration, little research has been done on identifying it. The majority of current research focuses on hate speech in general, which leaves a gap in addressing these particular and culturally relevant types of online abuse. In this work, we explore machine learning models to detect hate speech related to caste and migration in Tamil code-mixed text using TF-IDF and pretrained model embeddings to identify the most effective approaches for handling this task.

2 Data

The dataset used for this task includes text samples from social media platforms, including posts that are general and those that are specifically related to caste or migration hate speech, along with the labels that correspond to these posts for the purpose of identifying hate speech (Rajiakodi et al., 2025). The dataset includes three different language representations: English, Tamil, and Tanglish (a code-mixed Tamil and English). The provided train and development datasets were merged into a single training dataset. An overview of this combined dataset's distribution across classification labels can be found in Table 1.

Label	Count
Caste/Migration Hate Speech	2399
Not Caste/Migration Hate Speech	3900

Table 1: Dataset distribution across classification labels on train and development datasets combined

3 Methodology

This section describes the methodology followed which includes data pre-processing, feature extraction and model training used in this study. The codebase is available at our GitHub repository[2].

3.1 Data Preprocessing

A number of text preprocessing procedures were used to make sure the dataset was clean and appropriate for classification. Initially, the text's hashtags were taken out and processed independently. A word segmentation model was employed to separate hashtags into meaningful components because they frequently contain compound words without spaces. To preserve their semantic meaning, the processed hashtags were subsequently added back to the main body of text.

In order to anonymize the users tagged, while preserving the conversation's structure, user mentions (such as @username) were replaced with the placeholder <USER>. The emoji library was also used to translate emojis into their textual descriptions, guaranteeing that the text retained the emojis' sentiment and meaning. Lastly, to standardize the input format, extra whitespace and newline characters were eliminated. By removing noise from the text, these preprocessing techniques assisted in preserving the most important linguistic information.

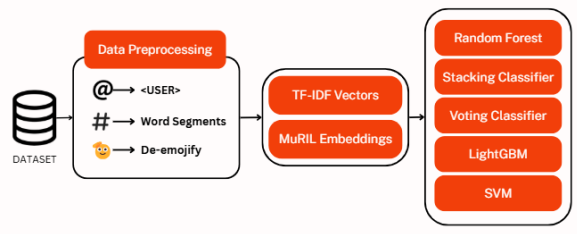

Figure 1: Methodology

3.2 Feature Extraction

Transformer-based embeddings and statistical methods were the two main strategies used for fea-

[2]https://github.com/amri-tah/ItsAllGoodMan-LT-EDI-2025

ture extraction for our task. Significant patterns in text have been captured using TF-IDF vectors which was used to train machine learning models. Furthermore, using the contextual understanding offered by these pre-trained language models, embeddings from mBERT, XLM-RoBERTa, and MuRIL were extracted and utilized as input features for machine learning models. To further investigate the effect of hybrid feature representations on classification performance, we experimented contatenating the most effective embeddings, TF-IDF and MuRIL.

3.3 Traditional Machine Learning Models

A range of machine learning models, including ensemble-based and individual classifiers, were investigated. The individual classifiers that were employed included XGBoost, logistic regression, decision trees, SVM, Random Forest, gradient boosting, and LightGBM.

A voting-based method aggregated three best performing classifiers, by averaging probability scores (soft voting) and by choosing the most often predicted class (hard voting). The strengths of several top-performing base models were combined using the stacking approach, and a logistic regression model was employed as the last decision layer.

3.4 Hyperparameter Tuning

Grid Search was used for hyperparameter tuning, in order to maximize the model performance. Various machine learning models such as random forest, logistic regression, etc, were tested with various learning rates, depth values, and weight adjustments. To evaluate model stability and make sure the models don't overfit, a 10-fold cross-validation technique was also applied.

4 Results

The results of several machine learning models on different text representations have been explored in this section. Initially, machine learning models were trained using TF-IDF vectors and embeddings from a number of pre-trained models, including mBERT, Tamil BERT, LaBSE, XLM-Roberta, and MuRIL (base and large). Of these, TF-IDF vectors and MuRIL large embeddings outperformed the others on the validation split. Following this, a combination TF-IDF vectors and MuRIL embeddings were used to to further improve accuracy and F1-scores.

Table 2 presents the classification performance of the best performing machine learning models using text representations: TF-IDF vectors, MuRIL embeddings, and their combination. The evaluation metrics considered are accuracy and macro F1-score, where higher values indicate better performance.

Model	Accuracy	Macro F1
TF-IDF Vectors		
Random Forest	**0.80**	**0.77**
Stacking Classifier	0.79	0.76
Voting Classifier (Soft)	0.77	0.73
MuRIL Embeddings		
Stacking Classifier	**0.78**	**0.75**
XGBoost	0.77	0.74
Voting Classifier	0.77	0.73
TF-IDF + MuRIL Embeddings		
Voting Classifier (Soft)	**0.79**	**0.77**
XGBoost	0.78	0.77
LightGBM	0.78	0.76

Table 2: ML Models for Each Embedding Type

The validation of our models trained were done using the 20% of the dataset provided to us for training. Using this validation set, Random Forest classifier outperformed the other models trained on TF-IDF vectors, achieving a macro F1-score of 0.77 and an accuracy of 0.80, whereas Stacking Classifier performed the best for models trained on MuRIL embeddings, with a macro F1-score of 0.75 and an accuracy of 0.78. Voting classifier and XGBoost trained on MuRIL embeddings gave similar results with an accuracy of 0.77 and F1-scores of 0.73 and 0.74, respectively. On combining MuRIL representations with TF-IDF vectors, an overall improvement in classification performance can be observed.

Overall classification performance was improved by combining MuRIL based embeddings with TF-IDF vectors and training it on Soft Voting Classifier with an accuracy score of 0.79 and F1 score of 0.77. Following closely behind, XGBoost and LightGBM returned comparable results with an accuracy of 0.78 and F1 scores of 0.77 and 0.76 respectively. These findings imply that improving classification performance requires using both contextual embeddings and traditional statistical features. When combined with TF-IDF, MuRIL embeddings helped to improve performance, but they did not outperform TF-IDF-based models on

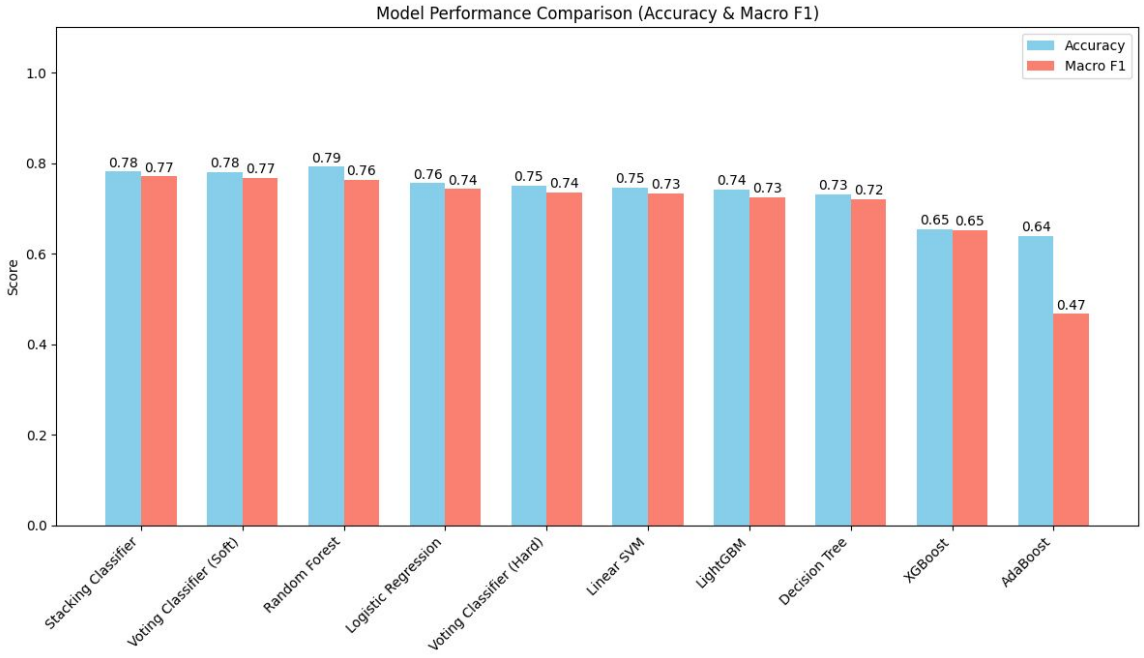

Figure 2: MuRIL + TF-IDF Model Training

their own. Overall, these results show how well hybrid feature representations work to produce reliable text classification outcomes.

In comparison to the best individual feature-based models (TF-IDF only and MuRIL only), the combination of MuRIL embeddings and TF-IDF vectors performed better on the held-out test set, obtaining the highest macro F1-score of 0.802. This supports how well contextual and statistical text representations work together to classify hate speech.

The term importance based on the frequency of explicit hateful words and keywords is useful in identifying whether a speech is hateful or not and is obtained by using statistical features such as TF-IDF. However, this feature alone might fail in some cases since it does not understand the semantic meaning behind these words, especially in code-mixed and multilingual contexts. This is where MuRIL embeddings comes into play, which, when combined with TF-IDF, has proven to be a good feature representation for the hate speech classification task.

5 Conclusion

This paper presents our system for the LT-EDI@LDK 2025 Shared Task on detecting caste and migration hate speech across Tamil, Tanglish, and English. We experimented with both TF-IDF vectors and transformer embeddings (especially MuRIL) as input features for a range of machine learning classifiers.

Our experiments clearly showed that combining traditional TF-IDF vectors with the contextual understanding from MuRIL embeddings produced the best outcome. Specifically, a Soft Voting Classifier using this hybrid TF-IDF + MuRIL feature set achieved the highest macro F1-score of 0.802 on the competition's test data. Using both TF-IDF and MuRIL together produced a better score than using either one individually. This likely happened because the two methods capture different kinds of useful information. TF-IDF finds key hate terms through frequency, while MuRIL understands the context and nuance, essential for the code-mixed and multilingual text we analyzed.

Our system using this method placed 6th overall in the shared task. This work shows that blending statistical text features with modern contextual embeddings offers a solid path forward for effectively detecting hate speech in complex, real-world linguistic scenarios like those found in Indian social media.

6 Limitations

Our model performances have been primarily validated on the provided LT-EDI@LDK 2025 dataset, therefore the generalization of the models on the full diversity of online caste and migration hate speech might be constrained.

References

Md Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_Manning@LT-EDI 2024: Transformer-based approach on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 238–243, St. Julian's, Malta. Association for Computational Linguistics.

Shibly FHA, Uzzal Sharma, and HMM. Naleer. 2023. Development of an efficient method to detect mixed social media data with tamil-english code using machine learning techniques. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

K Jaidev, Munnangi Pranish Kumar, Jampala Sai Chandana, Charishma Chowdary, and Sachin Kumar. EasyChair, 2024. Offensive text detection: Exploring traditional classifiers, ensemble models, and kolmogorov arnold networks in code-mixed tamil-english text. EasyChair Preprint 15581.

Sreelakshmi K, Premjith B, and Soman Kp. 2021. Amrita_CEN_NLP@DravidianLangTech-EACL2021: Deep learning-based offensive language identification in Malayalam, Tamil and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, Kyiv. Association for Computational Linguistics.

Jayanth Mohan, Spandana Reddy Mekapati, Premjith B, Jyothish Lal G, and Bharathi Raja Chakravarthi. 2025. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. volume 24, New York, NY, USA. Association for Computing Machinery.

Jakub Pokrywka and Krzysztof Jassem. 2024. kubapok@LT-EDI 2024: Evaluating transformer models for hate speech detection in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 196–199, St. Julian's, Malta. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Nishanth S, Shruthi Rengarajan, S Ananthasivan, Burugu Rahul, and Sachin Kumar S. 2025. ANSR@DravidianLangTech 2025: Detection of abusive Tamil and Malayalam text targeting women on social media using RoBERTa and XGBoost. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 711–715, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Dhyan Singh. 2025. Dalits' encounters with casteism on social media: a thematic analysis. volume 28, pages 335–353. Routledge.

Kriti Singhal and Jatin Bedi. 2024. Transformers@LT-EDI-EACL2024: Caste and migration hate speech detection in Tamil using ensembling on transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253, St. Julian's, Malta. Association for Computational Linguistics.

NSR@LT-EDI-2025: Automatic speech recognition in Tamil

Nishanth S, Shruthi Rengarajan, Burugu Rahul, G. Jyothish Lal
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22149, cb.en.u4aie22154,
cb.en.u4aie22161}@cb.students.amrita.edu
g_jyothishlal@cb.amrita.edu

Abstract

Automatic Speech Recognition (ASR) technology can potentially make marginalized communities more accessible. However, older adults and transgender speakers are usually highly disadvantaged in accessing valuable services due to low digital literacy and social biases. In Tamil-speaking regions, these are further compounded by the inability of ASR models to address their unique speech types, accents, and spontaneous speaking styles. To bridge this gap, the LT-EDI-2025 shared task is designed to develop robust ASR systems for Tamil speech from vulnerable populations. Using whisper-based models, this task is designed to improve recognition rates in speech data collected from older adults and transgender speakers in naturalistic settings such as banks, hospitals and public offices. By bridging the linguistic heterogeneity and acoustic variability among this underrepresented population, the shared task is designed to develop inclusive AI solutions that break communication barriers and empower vulnerable populations in Tamil Nadu.

Keywords: Speech Recognition, Indian languages, Tamil, Whisper model

1 Introduction

Speech is the most natural form of human communication. As Speech Technologies and AI advance rapidly, Automatic Speech Recognition (ASR) systems have become essential for human-computer interactions, offering convenience especially in multilingual countries like India. However, a significant gap exists in how these technologies reach vulnerable communities.But there is a big disconnect in how these technologies get to communities that are at risk.

Elderly people and transgender communities are marginalized groups that struggle to access essential services in Tamil-speaking areas. While transgender communities, who are frequently the targets of discrimination in society, encounter obstacles in education and digital literacy, older adults struggle with age-related disabilities that make interacting with digital interfaces more difficult. The most dependable method of expressing needs for both groups is still face-to-face interaction; however, current ASR systems that were trained on mainstream data are unable to accurately transcribe their distinct speech patterns.

Developing reliable ASR systems for these groups involves multiple challenges. Elderly speakers show acoustic variability due to physiological changes like decreased vocal cord elasticity and altered speech rhythm. Transgender speakers may present varied vocal characteristics influenced by hormonal treatments or voice therapy. Tamil itself poses linguistic challenges as a Dravidian language with complex morphology, multiple dialects, and code-mixing tendencies.

Compounding these issues is the severe shortage of high-quality speech corpora from older and transgender Tamil speakers. Available resources typically focus on standard varieties with minimal coverage of marginalized voices.

To address these gaps, the LT-EDI 2025 Shared Task on Speech Recognition for Vulnerable Speakers in Tamil(B. Bharathi, 2025) aims to stimulate research in this underexplored area. The task focuses on developing ASR systems to transcribe spontaneous Tamil speech from older and transgender speakers, providing curated datasets and leveraging models like Whisper to tackle the unique challenges these populations present.

This collaborative initiative primarily seeks to bridge the technological divide and enhance digital inclusion for underserved communities, empowering vulnerable Tamil speakers with improved access to essential services through accurate speech recognition.

More details about the shared task can be found

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 95–99

at[1].

2 Related works

Recent work in ASR has gained interest in calibrating systems to function proficiently with underrepresented and non-standard speech categories. (S and B, 2022) introduced a Transformer-based Tamil conversational ASR for the speech of older and trans women. Training their system with actual audio captured from public spaces, they confronted natural speech variations and obtained a WER of 39.65%, divulging the sophistication in creating accessible ASR. In a similar vein, (R et al., 2024) developed ASR for vulnerable Tamil speakers based on fine-tuned Whisper and XLS-R models. Fine-tuned on LT-EDI@EACL2024 data, the Whisper model performed with improved robustness with a WER of 24.452, proving its ability to deal with age, gender, and social background-caused variability.

(Radford et al., 2022) presented a large-scale method by training ASR models on 680,000 hours of weakly labeled web audio. Their zero-shot models saw success across several benchmarks and demonstrated nearly human-level performance and extreme generalizability over languages and tasks. (Shraddha et al., 2022), in contrast, investigated ASR performance on child speech, a field usually overlooked because of the scarcity of data and pitch/articulation contrasts. Benchmarking six end-to-end models, they highlighted the limitation of adult-trained ASR systems when applied to child speech.

(Biswas et al., 2022) suggested the application of Weighted Finite-State Transducers (WFSTs) to integrate acoustic, lexical, and language models within a probabilistic and efficient decoding pipeline. As opposed to end-to-end neural systems, WFSTs are modular and flexible and thus effective for structured ASR frameworks. Overall, these studies highlight the requirement of ASR systems that are robust, flexible, and inclusive, capable of accommodating linguistic, acoustic, and demographic variation.

3 Dataset

The dataset was distributed by the shared task organisers of Speech Recognition for Vulnerable Individuals in Tamil - LT-EDI@LDK 2025 (B et al., 2022)(B et al., 2024). The audio samples are collected from individuals whose mother tongue is Tamil and was presented in a .wav format. The total audio length is 7 hours and 30 minutes and is divided into 5.5 hours(approx.)

4 Methodology

A speech recognition model (Whisper-V3-large) (Graham and Roll, 2024) is trained using a specially created audio-text dataset. The suggested methodology makes use of two folders, one of which contains audio recordings and the other of which contains the text transcripts that go with them. Every audio file has a corresponding text file with the same name but a different extension. First, each transcript's encoding is automatically identified to guarantee accurate text file reading. Because text files can be saved in a variety of formats and incorrect reading can result in errors or misinterpreted characters, this step is crucial. After being identified, the transcript is read and cleared of any extraneous words before being paired with the matching audio file in a structured list.

Following the collection of all legitimate audio-transcript pairs, the information is transformed into a specific format that facilitates effective audio data handling. The way the audio column is handled enables the system to directly load and process waveform data, which makes it appropriate for speech recognition model training. The dataset is divided into two sections, usually in a 90-10 ratio, one for training and one for testing, in order to accurately assess model performance.

Other components are initialized to get this data ready for model training. These comprise tools for converting text into numerical form, extracting significant features from unprocessed audio, and combining the two to expedite input processing. In addition, a pre-trained model architecture is loaded, which is intended to convert spoken language into written formis loaded. All of these elements are set up to function exclusively with a selected language, in this example Tamil, guaranteeing that the input and output match the features of that language. When combined, the dataset and these tools form a comprehensive pipeline that is prepared for training a model to comprehend and record spoken Tamil.

[1] https://codalab.lisn.upsaclay.fr/competitions/21879

4.1 Data Preprocessing

To be able to train a model that translates spoken language into written text, the suggested methodology first entails the preparation and compilation of audio-text data. Each data sample is first processed, with audio signals being converted into model-appropriate numerical features and the corresponding transcriptions being changed into token identifier sequences that the model can comprehend. To make sure that the input and output are in a format that the model can use for learning, this transformation is carried out for the entire dataset. A distinctive component is defined to appropriately arrange and align the inputs and outputs in order to manage training in batches. Since text labels and audio inputs need different handling, they are separated. To enable effective computation, the audio features are gathered and padded in a batch so that they are all the same size. In a similar manner, the text labels are padded to guarantee size alignment, but with special handling: portions of the padding are designated to be disregarded during training to prevent the model from treating them as actual words. Additionally, since the training process will add the starting tokens separately, any extraneous ones that are already in the labels are eliminated to prevent duplication. By ensuring that the model receives clean, consistent, and well-aligned data at every stage, this meticulous preparation raises the training process's accuracy and efficiency.

4.2 Training and Model Evaluation

A training framework for a sequence-to-sequence model by defining key components like training parameters, datasets for training and evaluation, and a data preparation function is set up. It contains a preprocessor to format model input data and a metric calculation function to track performance. By automating the training and evaluation process, this setup makes it possible to fine-tune the model effectively.

The parameters for training a sequence-to-sequence model are established by the given configuration. It outlines specifics such as the model's storage location, the training and evaluation batch sizes, the number of training epochs, and the frequency of evaluations and checkpoint saves. To handle large models, the training process is optimized using techniques like gradient accumulation, mixed-precision training, and memory-efficient methods. The model that performs the best is saved for later use after its performance is assessed on a regular basis using a particular metric. Furthermore, the generated sequences are constrained to a specific length during prediction, and the learning rate is initially increased gradually.

The two main metrics derived for evaluation are the Word Error Rate (WER) and the Character Error Rate (CER), both of which are used in common practice to measure the precision of automatic speech recognition (ASR) systems (Hamed et al., 2023). WER calculates the number of total errors in the transcription of the model against the ground truth set. This metric is calculated by comparing the predicted transcription with the reference text and penalizing insertions, deletions, and word substitutions. The lower the WER, the better the performance. CER is similar in operation but at the character level rather than the word level. CER calculates the number of errors at the character level, which can be useful when the model transcribes text with various spelling or formatting errors. Both WER and CER have significance since both provide a precise understanding of the accuracy of the model at varying levels of granularity (word vs. character).

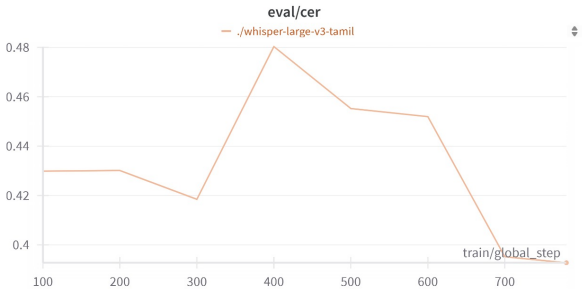

Figure 1: CER evaluation graph

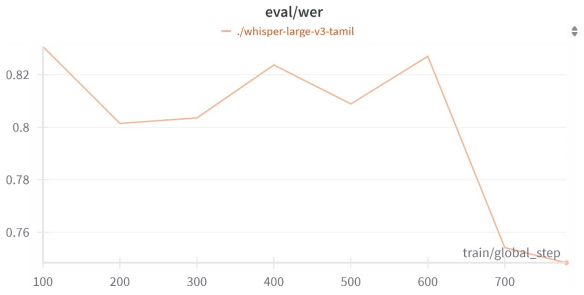

Figure 2: WER evaluation graph

Table 1: Step-wise Performance of the Whisper Model on the Tamil Dataset

Step	Training Loss	Validation Loss	WER ()	CER ()
100	0.3993	0.6537	0.8307	0.4299
200	0.1820	0.8540	0.8015	0.4301
300	0.1460	0.9717	0.8035	0.4185
400	0.1226	0.9849	0.8237	0.4803
500	0.1006	1.0915	0.8089	0.4552
600	0.0272	1.1300	0.8270	0.4519
700	0.0035	1.1713	0.7541	0.3952

Parameter	Value
output_dir	./whisper-large-v3-tamil
per_device_train_batch_size	16
gradient_accumulation_steps	2
learning_rate	3e-4
warmup_steps	500
num_train_epochs	30
gradient_checkpointing	True
fp16	True
evaluation_strategy	steps
per_device_eval_batch_size	8
predict_with_generate	True
generation_max_length	225
save_steps	500
eval_steps	100
logging_steps	50
report_to	wandb
load_best_model_at_end	True
metric_for_best_model	wer
greater_is_better	False
save_total_limit	2
push_to_hub	False
dataloader_num_workers	7
dataloader_prefetch_factor	2
dataloader_pin_memory	True

Table 2: Updated Training Parameters for Seq2Seq Model

Evaluation Type	Score
Word Error Rate (WER)	0.7484
Character Error Rate (CER)	0.3927

Table 3: Model Evaluation

5 Experimental Inference

In this proposed methodology, a pre-trained Whisper model from Hugging Face is leveraged to perform automatic speech recognition (ASR) (Amorese et al., 2023).

The script randomly selects a sample from the test set of a preloaded dataset. The sample contains an audio file and its corresponding transcribed text. The audio file is extracted from the sample, along with its sampling rate, which is important for the subsequent feature extraction process. The audio data, in the form of a raw waveform, is passed to the Whisper Processor, which is responsible for converting the raw audio into features that the Whisper model can understand. Then, the waveform is transformed into a spectrogram, a 2D representation of the audio, which is the input format the model expects.

Once the audio has been transformed into input features, it is sent to the model which takes the processed audio features as input and outputs a sequence of predicted token IDs that represent the transcription of the speech. These token IDs are converted from the numerical token IDs back into readable text, skipping any special tokens that may have been used during training (such as padding or end-of-sequence markers).

Finally, the original and machine-predicted scripts are compared, which allows an evaluation of the models performance, offering insight into how well the Whisper model has learned to transcribe speech. By printing both the original and predicted text, users can directly observe how accurately the model has transcribed the audio sample, which is the primary goal of the ASR process.

With the same method used to obtain inference on the test data, We secured **2nd** rank in this task.

The code files for this project can be accessed from[2]

6 Conclusion

This paper presented the results of the shared task addresses a challenging area in Automatic Speech Recognition: vulnerable old-aged and transgender people in Tamil. Many elderly people do not know how to use the equipment available to them. Speech is the only medium that could help transgender people meet their needs because they are denied access to primary education due to societal prejudice. Data on spontaneous speech are

[2] https://github.com/BURUGURAHUL/NSR-LT-EDI-2025-Automatic-speech-recognition-in-Tamil

collected from elderly and transgender individuals who cannot take advantage of these resources.

7 Limitations

While working on this topic, the major limitation we faced was the use of Whisper V3 large, which significantly increased computational requirements. Due to the models large size, standard GPUs were insufficient, and an NVIDIA A6000 was required to handle the memory load. This made the approach less accessible in environments with limited hardware resources.

References

Terry Amorese, Claudia Greco, Marialucia Cuciniello, Rosa Milo, Olga Sheveleva, and Neil Glackin. 2023. Automatic speech recognition (asr) with whisper: Testing performances in different languages. In *S3C@ CHItaly*, pages 1–8.

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya N, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Bharathi B, Bharathi Raja Chakravarthi, Sripirya N, Rajeswari Natarajan, Suhasini S, and Swetha Valli. 2024. Overview of the third shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

N. Sripriya Rajeswari Natarajan Rajalakshmi R S. Suhasini B. Bharathi, Bharathi Raja Chakravarthi. 2025. Overview of the Fifth Shared Task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Dipshikha Biswas, Suneel Nadipalli, B. Sneha, and M. Supriya. 2022. Speech recognition using weighted finite-state transducers. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–5.

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *The Journal of the Acoustical Society of America*, 4.

Injy Hamed, Amir Hussein, Oumnia Chellah, Shammur Chowdhury, Hamdy Mubarak, Sunayana Sitaram, Nizar Habash, and Ahmed Ali. 2023. Benchmarking evaluation metrics for code-switching automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 999–1005.

Jairam R, Jyothish G, Premjith B, and Viswa M. 2024. CEN_Amrita@LT-EDI 2024: A transformer based speech recognition system for vulnerable individuals in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 190–195, St. Julian's, Malta. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Suhasini S and Bharathi B. 2022. SUH_ASR@LT-EDI-ACL2022: Transformer based approach for speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 177–182, Dublin, Ireland. Association for Computational Linguistics.

S Shraddha, Jyothish Lal G, and Sachin Kumar S. 2022. Child speech recognition on end-to-end neural asr models. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6.

Solvers@LT-EDI-2025: Caste and Migration Hate Speech Detection in Tamil-English Code-Mixed Text

Mohanapriya K T[1], Anirudh Sriram K S[1], Devasri A[1], Bharath P[1], Ananthakumar S[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Hate speech detection in low-resource languages such as Tamil presents significant challenges due to linguistic complexity, limited annotated data, and the sociocultural sensitivity of the subject matter. This study focuses on identifying caste- and migration-related hate speech in Tamil social media texts, as part of the LT-EDI@LDK 2025 Shared Task. The dataset used consists of 5,512 training instances and 787 development instances, annotated for binary classification into caste/migration-related and non-caste/migration-related hate speech. We employ a range of models, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and transformer-based architectures such as BERT and multilingual BERT (mBERT). A central focus of this work is evaluating model performance using macro F1-score, which provides a balanced assessment across this imbalanced dataset. Experimental results demonstrate that transformer-based models, particularly mBERT, significantly outperform traditional approaches by effectively capturing the contextual and implicit nature of hate speech. This research underscores the importance of culturally informed NLP solutions for fostering safer online environments in underrepresented linguistic communities such as Tamil.

1 Introduction

Detecting hate speech in low-resource languages such as Tamil is a complex and crucial task, especially in the context of caste- and migration-related discrimination. Tamil, predominantly spoken in Tamil Nadu (India), Sri Lanka, and among diasporic communities, is rich in cultural and linguistic diversity, which poses unique challenges to Natural Language Processing (NLP). In recent years, the spread of hate speech targeting caste and migrant communities has escalated on social media platforms, demanding robust automatic detection systems that are socially aware and ethically grounded.

The identification of such harmful content is complicated by the nuanced ways in which caste and migration are discussed, often involving implicit language, sarcasm, and regional idioms. Additionally, Tamil's morphological richness, the scarcity of annotated corpora, and the limited availability of linguistic tools make it difficult to develop high-performance hate speech classifiers. To address these issues, the LT-EDI@LDK 2025 Shared Task released a manually annotated dataset in Tamil, categorizing instances as either caste/migration-related hate speech or non-caste/migration-related hate speech (Rajiakodi et al., 2024).

In this study, we evaluate the performance of several machine learning and deep learning models—specifically, Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and transformer-based architectures such as BERT and multilingual BERT (mBERT)—for the task of hate speech classification (Vaswani et al., 2017; Devlin et al., 2019). Given the class imbalance and the sensitive nature of the content, we use macro F1-score as the primary evaluation metric. Our findings show that transformer-based models, particularly mBERT, perform significantly better at capturing contextual cues and implicit hate. This work contributes to the growing body of research aimed at improving online safety in underrepresented linguistic communities and emphasizes the importance of culturally and ethically grounded NLP approaches to hate speech detection.

2 Literature Survey

Hate speech detection is an increasingly important task in natural language processing (NLP), particularly with the rise of social media platforms where offensive and abusive content is frequently encountered. While substantial progress has been made in hate speech detection for high-resource

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 100–104

languages like English, the task remains underdeveloped for low-resource languages such as Tamil. Early shared tasks like HASOC and OffensEval laid foundational work in this domain (Zampieri et al., 2019; Mandl et al., 2020). Tamil poses unique challenges due to its rich morphology, code-mixed usage with English, and the cultural sensitivity of topics like caste and migration. The scarcity of large, annotated datasets in Tamil further complicates model development for hate speech classification.

Recent shared tasks such as the LT-EDI (Language Technology for Equality, Diversity, and Inclusion) series have helped bring attention to the issue, offering benchmark datasets and encouraging the development of hate speech detection systems specifically for Tamil and other Dravidian languages (Rani et al., 2022). These initiatives have laid the groundwork for evaluating traditional machine learning, deep learning, and transformer-based approaches on nuanced categories such as caste and migration hate speech.

2.1 Hate Speech Detection in Tamil

Initial approaches to Tamil hate speech detection involved traditional machine learning techniques, with Support Vector Machines (SVM) and Naive Bayes being among the earliest models. These methods utilized hand-crafted features like n-grams, part-of-speech tags, and TF-IDF vectors. Although simple and interpretable, these models often struggled to capture the semantic depth and informal variations in Tamil text, especially in code-mixed settings.

Deep learning methods, such as Convolutional Neural Networks (CNNs), were later introduced to overcome the limitations of feature engineering. CNNs have proven effective in capturing local dependencies in text, particularly through their use of convolutional filters across embedding sequences. Their success in image recognition tasks translated well to text classification by modeling syntactic patterns in sentence fragments.

More recently, transformer-based models such as BERT and multilingual BERT (mBERT) have transformed the field of NLP. These models employ self-attention mechanisms to capture contextual semantics across entire sentences, making them well-suited for detecting implicit hate speech and nuanced expressions. For Tamil, mBERT's multilingual training across 104 languages has proven particularly effective in low-resource scenarios by transferring knowledge from related high-resource languages (Devlin et al., 2019).

Vasantharajan and Thayasivam (2021) demonstrated the effectiveness of mBERT in classifying Tamil code-mixed YouTube comments. Similarly, Benhur and Sivanraju (2021) applied mBERT to the Tanglish dataset, achieving competitive results. These studies highlight the advantage of using transformer-based models in multilingual and culturally diverse contexts.

2.2 Caste and Migration Hate Speech

Detecting hate speech related to caste and migration in Tamil presents unique challenges due to the deep cultural and historical roots of these social structures. The language used in such contexts often includes sarcasm, indirect references, and culturally embedded terms, which are difficult to classify using traditional models. Studies have shown that transformer models like mBERT and BERT are better suited to handle such implicit and contextual expressions of hate speech.

For instance, Alam et al. (2024) conducted a comparative analysis of various transformer-based models on caste- and migration-related Tamil hate speech data, concluding that mBERT consistently delivered the best performance, with a macro F1-score of 0.80. Their work validates the use of multilingual transformers for sensitive and domain-specific tasks, particularly in underrepresented languages like Tamil.

2.3 Challenges in Hate Speech Detection for Tamil

Despite progress, several challenges continue to hinder the development of robust hate speech detection systems for Tamil:

Data Scarcity: The lack of large-scale, annotated datasets tailored to caste and migration hate speech remains a major barrier.

Code-Switching: Frequent code-mixing between Tamil and English on social media creates ambiguity for monolingual models.

Cultural Nuance: Tamil expressions of hate often include regional dialects, idioms, and sarcasm that require culturally aware annotation and modeling.

Informality: The informal and noisy nature of social media content makes tokenization, POS tagging, and syntactic parsing more difficult.

These factors collectively call for the use of sophisticated models like BERT and mBERT, which

can encode complex context and benefit from multilingual pretraining. However, even these models are constrained by the quality and quantity of labeled data available for fine-tuning.

2.4 Transformer Models and Advances

Transformer models such as BERT, multilingual BERT (mBERT), and MuRIL have revolutionized NLP, particularly for tasks in low-resource settings. Introduced by (Vaswani et al., 2017), the Transformer architecture enables parallel processing and better captures long-range dependencies, which are essential for understanding context in hate speech.

In the context of Tamil, mBERT and MuRIL have shown superior results due to their multilingual training on large-scale corpora. These models can transfer knowledge from high-resource languages to Tamil, thereby compensating for the lack of labeled data. IndicBERT, which is trained on 12 Indian languages, including Tamil, has also been explored for its lightweight architecture and adaptability to resource-constrained environments (Kakwani et al., 2020).

Despite these innovations, there remains a pressing need for more annotated datasets, domain-specific pretraining, and culturally sensitive modeling approaches (Hendrycks et al., 2021; Touvron et al., 2023) to further improve hate speech detection systems in Tamil.

3 Materials and Methods

3.1 Dataset Description

These texts are code-mixed with English and centered around themes of caste and migration, annotated for hate speech detection (Rajiakodi et al., 2024). The dataset is divided into three subsets: a training set with 8,042 samples, a validation set (dev.csv) with 1,006 samples, and a test set (test.csv) containing 1,001 samples. Each sample is labeled as either "Hate Speech," "Non-Hate Speech," or "Offensive but not Hate," allowing multi-class classification.

The class distribution in the training set is imbalanced, with "Non-Hate Speech" forming the majority, followed by a smaller proportion of "Hate Speech" and "Offensive" samples. To handle this imbalance, oversampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and random duplication were employed during training. This annotation and structure provide a nuanced foundation for detecting subtle and explicit hate expressions related to caste and migration.

3.2 Pre-processing and Feature Extraction

Due to the informal and code-mixed nature of the dataset, preprocessing was a critical step to improve model performance. The following techniques were employed:

Text Normalization: All text was lowercased, and punctuation, special characters, and elongated words were normalized. Hashtags were split when meaningful.

Tokenization: Tokenization was performed using Hugging Face tokenizers for BERT and mBERT, while standard NLP tokenizers were used for SVM and CNN.

Noise Removal: URLs, mentions (@username), emojis, and numbers were removed to reduce noise in social media-style texts.

Code-Mixing Handling: To address Tamil-English mixed content, language identification was applied. When possible, transliteration was used to normalize Tamil content written in Roman script.

Stopword Removal: Tamil and English stopword lists were used to eliminate non-informative tokens, mainly for traditional feature-based models like SVM.

Feature Extraction varied depending on the model type:

SVM: Employed Bag-of-Words (BoW) and TF-IDF vector representations to convert text into numerical features.

CNN: Used pretrained FastText Tamil word embeddings (Bojanowski et al., 2017) to capture local semantic and syntactic patterns.

BERT and mBERT: Fine-tuned transformer models with contextual embeddings that capture deep semantic features. Sentence embeddings generated by these models provided rich context representations, crucial for identifying implicit hate speech. Prior research has shown the utility of such embeddings in low-resource settings (Bouraoui et al., 2020).

3.3 Proposed Classifiers

Four classifiers were developed and evaluated, categorized as follows:

Traditional Model:

Support Vector Machine (SVM): Selected for its robustness in high-dimensional spaces and interpretability. TF-IDF features provided the best

performance among traditional vector representations.

Deep Learning Model:

Convolutional Neural Network (CNN): Designed to learn local n-gram features from FastText embeddings. It effectively captured spatial hierarchies in the text data.

Transformer-Based Models:

BERT: Fine-tuned on the hate speech dataset to leverage deep contextual embeddings.

Multilingual BERT (mBERT): Trained on a large multilingual corpus, mBERT was particularly effective in handling code-mixed Tamil-English text and showed robust generalization across hate speech categories.

4 Results and Discussion

The hate speech detection experiments using Tamil-English code-mixed data demonstrate that transformer-based models, particularly Multilingual BERT (mBERT), achieve the highest performance compared to traditional and shallow learning models. mBERT excels in capturing complex, context-rich semantic structures within code-switched and informal text. Deep learning models such as CNN and BiLSTM performed moderately well, capturing local and sequential patterns, respectively. Traditional models such as SVM, Logistic Regression, and KNN were able to handle basic binary discrimination but struggled with fine-grained category separation.

4.1 Performance Metrics

All models were evaluated on the development set using standard metrics: Accuracy, Precision, Recall, F1-score, and Macro-Averaged scores to ensure balance across class imbalances. The confusion matrices revealed that the major source of error involved misclassifications between the *Hate Speech* and *Offensive* classes, suggesting semantic overlap and subtle contextual differences.

Table 1: Performance Comparison of Top Models

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)
CNN	79.0	77.5	74.7	75.9
BERT	80.0	79.0	78.0	78.0
mBERT	**80.0**	**80.0**	**78.0**	**79.0**

Among the tested models, mBERT consistently achieved the best macro-averaged F1-score (79%), indicating reliable performance across both majority and minority classes. Its ability to handle code-mixed inputs effectively distinguishes it in multilingual contexts. Compared to BERT, mBERT displayed marginal but consistent improvements in recall and macro-F1, especially for underrepresented hate-related classes.

4.2 Limitations

Despite promising results, several limitations were encountered:

Class Imbalance: The dataset exhibits a skewed distribution with fewer samples in the "Hate Speech" and "Offensive" categories. This imbalance led to minor bias in model predictions toward the majority class. Though techniques like SMOTE were employed, they couldn't fully replicate the diversity and complexity of real hateful content. Research shows that data augmentation using back-translation or paraphrasing can enhance minority class learning (Barro et al., 2023).

Computational Cost: Transformer-based models like mBERT require significant computational resources and training time. This makes real-time or edge deployment challenging. Future work may focus on model distillation or lighter variants like DistilBERT or ALBERT to balance performance and efficiency.

Code-Mixing Complexity: Many samples contain informal Tamil-English blends or slang that are difficult to tokenize or embed correctly. This causes confusion particularly between offensive and hate speech categories. More robust language identification and translation pipelines may mitigate this issue.

Limited Annotated Data: The lack of large-scale annotated code-mixed Tamil datasets restricts model generalization. Introducing active learning and human-in-the-loop feedback mechanisms could help create a continuously evolving and more representative dataset.

5 Conclusion

In this study, we addressed the challenge of hate speech detection in Tamil-English code-mixed social media text, a complex task due to linguistic diversity, informal language, and limited annotated resources. We implemented and evaluated a classification pipeline using a combination of traditional and modern techniques, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and transformer-based models such

as BERT and multilingual BERT (mBERT).

Our experiments demonstrated that transformer-based models, particularly mBERT, achieved the best overall performance, effectively capturing contextual and semantic nuances in code-mixed text. CNNs provided competitive results by modeling local syntactic patterns, while SVM served as a strong baseline for traditional feature-based learning in low-resource conditions.

Despite the encouraging results, several challenges remain unresolved, including class imbalance, informal code-switching, and the limited size of annotated datasets. Future work may focus on incorporating domain-adaptive pretraining, leveraging data augmentation strategies, and employing multilingual knowledge integration to further improve model robustness. This research contributes to the development of effective NLP solutions for underrepresented South Asian languages and supports broader efforts to ensure safer, more inclusive online spaces.

6 Project Repository

The full source code for this project is available on GitHub: Bharath

References

Sarah Barro, Marcos Zampieri, and Ahmed Abdelali. 2023. Investigating the impact of data augmentation techniques for low-resource hate speech detection. In *Proceedings of ACL 2023*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. In *Proceedings of AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Dan Hendrycks, Collin Burns, Steven Basart, and et al. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Divyanshu Kakwani, Raghav Aggarwal, Siddhant Garg, and et al. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. *Findings of EMNLP*.

Thomas Mandl, Sandip Modha, Pooja Rani, and et al. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages. In *Working Notes of FIRE 2020*.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Pooja Rani, Thomas Mandl, Sandip Modha, and et al. 2022. Hasoc 2022: Hate speech and offensive content identification in indic languages. In *Working Notes of FIRE 2022*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, and et al. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*, pages 1415–1420.

CUET_N317@LT-EDI 2025: Detecting Hate Speech Related to Caste and Migration with Transformer Models

Md. Nur Siddik Ruman, Md. Tahfim Juwel Chowdhury, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u2004098, u2004094}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Language that criticizes, threatens, or discriminates against people or groups because of their caste, social rank, or status is known as caste and migration hate speech, and it has grown incredibly common on social media. Such speech not only contributes to social disruption and inequity, but it also puts at risk the safety and mental health of the targeted groups. Due to the absence of labeled data, the subtlety of culturally unique insults, and the lack of strong linguistic resources for deep text recognition, it is especially difficult to detect caste and migration hate speech in low-resource Dravidian languages like Tamil. In this work, we address the Caste and Migration Hate Speech Detection task, aiming to automatically classify user-generated content as either hateful or non-hateful. We evaluate a range of approaches, including a traditional TF-IDF-based machine learning pipeline using SVM and logistic regression, alongside five transformer-based models: mBERT, XLM-R, MuRIL, Tamil-BERT, and Tamilhate-BERT. Among these, the domain-adapted Tamilhate-BERT achieved the highest macro-F1 score of 0.88 on the test data, securing 1st place in the Shared Task on Caste and Migration Hate Speech Detection at DravidianLangTech@LT-EDI 2025. Our findings highlight the strong performance of transformer models, particularly those fine-tuned on domain-specific data, in detecting nuanced hate speech in low-resource, code-mixed languages like Tamil.

1 Introduction

Caste and migration related hate speech is defined as language that insults, threatens, or discriminates against individuals or groups based on their caste, social status, or immigration background, has be- come increasingly prevalent on social media (Gagliardone et al., 2015). This type of speech affects mental health in a very bad way. So, detection of caste- and migration-related hate speech is very crucial. The automatic hate speech tool may be useful to prevent such activities. Detecting hate speech in a low-resource lan- guage like Tamil is very challenging due to lim- ited resources. In addition, our task was to identify only caste- and migration-related hate speech. To find out whether a sentence expresses hate or not it is crucial to understand the intent of the language (Schmidt and Wiegand, 2017). In Tamil, it is quite difficult to identify compared to high-resource language like English. In this study, we have fine- tuned several models to facilitate automatic hate speech detection in Tamil.

1. Proposed a Transformer based model with ensembles

2. Analyzed several ML and Transformer-based models for detecting hate speech in Tamil

Our code is available in this link GitHub repository.

2 Related Work

There have several research recently in identifying hate speech related to caste and migration in Tamil. Transformer models (Vaswani et al., 2017) have become foundational, with several recent studies exploring their efficacy in this specific domain. For the LT-EDI 2024 shared task on Caste and Migration Hate Speech Detection in Tamil (Rajiakodi et al.), Alam et al. (2024) investigated various models, finding M-BERT to achieve a macro F1-score of 0.80. In the same shared task, Singhal and Bedi (2024) demonstrated the power of ensembling transformer-based models (XLM-R, mBERT, MuRIL) through majority voting, securing the 1st rank with a macro F1-score of 0.82.

Beyond monolingual text, Hossain et al. (2025a) explored multimodal fusion for Telugu hate speech, also reporting a strong unimodal text baseline with

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 105–110

mBERT (F1-score 0.4968). The challenge of code-mixed Dravidian languages was addressed by Sreelakshmi et al. (2024), who found a combination of MuRIL embeddings and an SVM classifier to be highly effective, achieving accuracies up to 96%. For Indonesian, another context outside of high-resource languages, Hakim et al. (2024) combined IndoBERTweet with BiLSTM and CNN, yielding an F1-score of 85.06%.

Ensemble strategies remain popular; Roy et al. (2022) proposed a weighted ensemble of BERT models and a deep neural network for offensive and hate speech in Tamil and Malayalam code-mixed data, achieving high F1-scores. Highlighting the challenges in low-resource settings, Reddy et al. (2024) investigated data augmentation and noted the ineffectiveness of POS tagging for Dravidian languages. Multimodal approaches have also been investigated by Hossain et al. (2025b), who propose a transformer-based multimodal fusion model with cross-modal attention for hate speech detection.

3 Dataset and Task Description

The dataset (Ponnusamy et al., 2024) for the Caste and Migration Hate Speech detection task consists of mixed Tamil-English social media comments that have been annotated to indicate whether hate speech related to caste or migration is present (1) or not (0). Three separate sets of data are offered: training, development, and test.

The class-wise distribution of the dataset is summed up in Table 1. The class imbalance in both splits is similar, with approximately 62% of cases falling into the No Hate Speech class. An important factor in our modeling strategy is this imbalance.

Set	Non Hate Speech (0)	Hate Speech (1)	Total
Train	3,415 (61.96%)	2,097 (38.04%)	5,512
Development	485 (61.63%)	302 (38.37%)	787
Test	970 (61.55%)	606 (38.45%)	1,576

Table 1: Class-wise distribution of the dataset.

4 System Overview

Text classification tasks are currently very challenging for low-resource languages in social media like Tamil. Therefore, to detect Tamil hate speech related to caste and migration, we followed two distinct paths: a traditional machine learning approach as a baseline and advanced transformer-based approaches to handle the complexities of the text corpus. The figure depicts the overall process flow that we applied to do the task.

4.1 Data Preprocessing

This work is part of the Caste and Migration Hate Speech Detection shared task (Rajiakodi et al., 2025). The text corpus contained emojis, URLs, and mixed scripts, so we needed a pre-processing pipeline to manage it while preserving meaning. As a result, we performed text normalization to convert the whole corpus to lowercase to avoid case-sensitive mismatches. We applied regular expressions to strip out mentions like @username, URLs, numbers, and special characters. We also kept Tamil script (Unicode range `\u0B80-\u0BFF`) and English letters for code-mixed text. For the ML approach, we used the IndicNLP tokenizer to break Tamil text into meaningful units and removed Tamil stopwords using a curated list from GitHub. For transformers, we applied model-specific tokenizers from HuggingFace.

4.2 Feature Extraction

Feature Extraction is the process to convert raw text into a format comprehensible to our models. We used different tactics related to each approach:

- **Traditional ML:** We applied TF-IDF vectorization after trying out simpler bag-of-words models. We looked at unigrams, bigrams, and trigrams to get the short phrases and restricted the functionality to 7,000 to maintain things manageable without compromising on essential patterns.

- **Transformer-Based Models:** In this, we allowed the models' pre-trained tokenizers to do the job of converting text to token IDs with a maximum of 256 tokens. We truncated longer texts and padded shorter ones.

4.3 Traditional ML Approach

For our first step in the problem task, we used an ensemble of ML classifiers on the train dataset to figure out the complexities in more broader aspect. We trained a support vector machine (SVM) and a Logistic Regression model. For SVM, we used a linear kernel and enabled probability outputs, which we found necessary for ensemble voting. For Logistic regression, we kept it straightforward with the default configuration, mostly relying on its

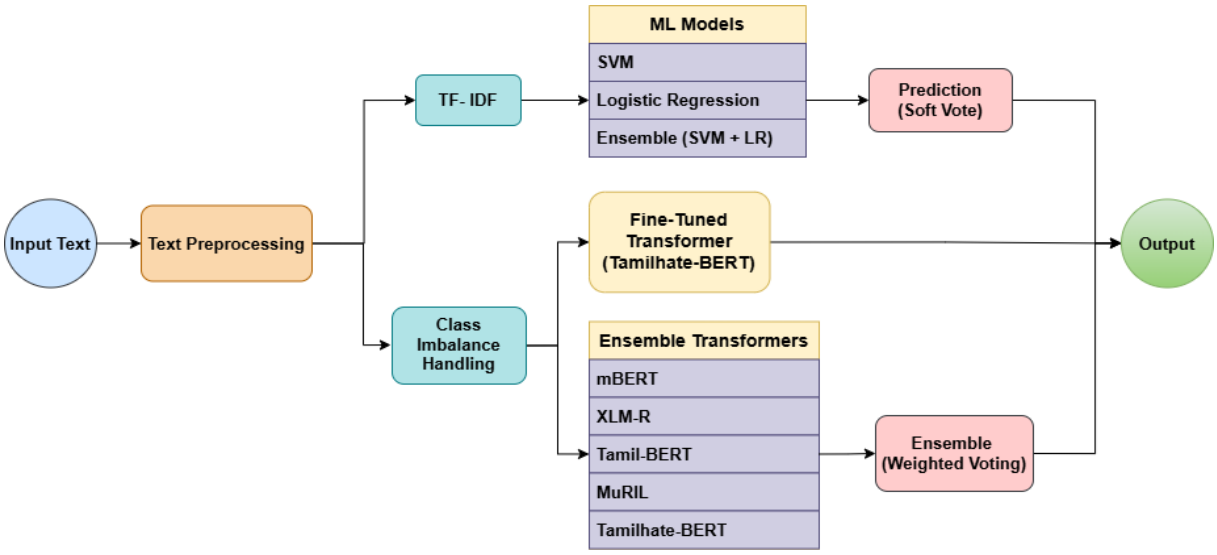

Figure 1: Schematic process of caste and migration hate speech detection in Tamil

probabilistic nature to complement the SVM. Then we combined these using a soft Voting Classifier, which averaged their predicted probabilities.

4.4 Transformer-Based Approach

Transformers are highly effective for understanding complex linguistic patterns and contextual relationships in text, especially in low-resource languages like Tamil. Therefore, we applied two strategies: a single fine-tuned model and a multi-model ensemble.

4.4.1 Fine-tuned Model

In this approach, we used a single fine-tuned transformer model to establish a strong, task-specific baseline by leveraging domain-adapted language understanding. We researched and found Tamilhate-BERT(mdo) model, a fine-tuned version of Tamil-BERT(Joshi, 2022) model on caste hate speech, which seemed like a perfect fit. To train this model, we used the AdamW optimizer with a learning rate of 1×10^{-5}, a batch size of 16, and trained for up to 10 epochs. We adopted early stopping after 2 epochs according to the validation F1 score, which saved us from overfitting. The dataset was imbalanced, so we calculated class weights based on inverse class frequencies and used them in a custom CrossEntropyLoss function. We logged metrics every 50 steps and kept the best model checkpoint based on macro F1.

4.4.2 Ensemble of Transformers

We combined multiple transformer models, hoping their diversity would make the system more robust. We fine-tuned five models: Tamil-BERT(Joshi, 2022), Tamilhate-BERT(mdo), MuRIL(Khanuja et al., 2021), mBERT(Devlin et al., 2019), and XLM-R(Conneau et al., 2020), each chosen for its strength in Tamil or multilingual aspects. Every model was fine-tuned with tailored hyperparameters, learning rates from 1×10^{-5} to 3×10^{-5} with batch sizes of 16 or 32 and trained for 12 to 14 epochs with early stopping after 3 epochs. We used class weights and gradient accumulation for some models to handle memory constraints. After that, we tested three different methods to combine predictions: majority voting, averaging probabilities, and weighted voting (using log-transformed validation F1 scores as weights). Weighted voting proved most effective after some experimentation.

Method	Classifier	Precision	Recall	Macro F1
ML	SVM	0.73	0.69	0.70
	Logistic Regression	0.72	0.65	0.65
	Ensemble	0.72	0.68	0.69
Transformers	mBERT	0.80	0.78	0.79
	XLM-R	0.80	0.77	0.78
	Tamil-BERT	0.79	0.78	0.78
	MuRIL	0.80	0.78	0.78
	Tamilhate-BERT	**0.88**	**0.88**	**0.88**
	Ensemble	0.84	0.81	0.82

Table 2: Performance of different systems on the test dataset.

5 Result and Analysis

Table 2 demonstrates the evaluation results of ML and Transformer models on the test set. Performance of the models was determined by the macro F1 score. The traditional machine learning technique with ensembling and soft voting achieved a macro F1 score of 0.69, reflecting a decent baseline but struggling to fully capture the linguistic complexity. Among the transformer-based models, XLM-R, Tamil-BERT and MuRIL each achieved macro F1 of 0.78 and the single fine-tuned Tamilhate-BERT achieved the highest macro F1 score of 0.88. This result highlights how effective domain-specific transfer learning can be. We have also explored an ensemble of multiple transformers, including mBERT, XLM-R, MuRIL, Tamil-BERT, and Tamilhate-BERT which unexpectedly reached a lower macro F1 score of 0.82. While ensembling added robustness, a well-targeted, fine-tuned model outperformed all others.

6 Error Analysis

We conducted a detailed error analysis to better understand the strengths and limitations of our best-performing model.

6.1 Quantitative Analysis

Figure 2 illustrates the performance of the top-performing model using a confusion matrix. The model correctly classified 876 out of 970 non-hate samples and 522 out of 606 hate speech samples. However, it misclassified 94 non-hate instances as hate and 84 hate instances as non-hate. The results reveal a slight bias toward the majority class (non-hate). While class weighting mitigated some imbalance, linguistic nuances in Tamil social media text, such as code-mixing, sarcasm, or context-dependent phrases, likely contributed to errors.

6.2 Qualitative Analysis

Figure 3 shows a few sample predictions of the best model on the test dataset. Some of the errors show that the model struggles with the informal or mixed language text. For example, one misclassification occurred where the text had clear caste-based insults written in mixed language. Another example written in Tamil was sarcastic and subtle, which was challenging for the model to interpret correctly. On the other hand, the model correctly classified some Tamil-English mixed texts properly.

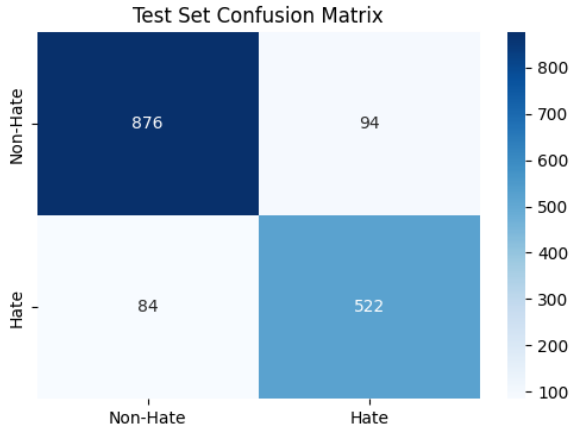

Figure 2: Confusion Matrix

Sample Text	Actual Label	Predicted Label
Holly பண்டிகை க்கு ஊருக்கு போரானுங்க திரும்பி கண்டீப்பா வருவானுங்க... (They go to their hometown for the Holi festival and will definitely come back.)	1	0
நாகப்பதனியா,நாகப்ப்பதனியா யார் பெரியவர் என்று பார்த்துவிடுவோம்… (Let's see who is greater, Nagapathan or Nagappathan.)	1	0
Nandu kadha theriyuma ? Tamizh nandu matum dhan mela yera vidama keezha thalite irundhudu.dats true (Do you know the crab story? Only Tamil crabs pull others down instead of letting them climb up. That's true.)	1	1
Avanunga holi festival ku poraanga yaa😂 (They're going for the Holi festival, yaa 😂)	1	1
தமிழர்களிடம் மட்டுமே வரவு செலவு வைத்துக் கொள்ள வேண்டும்! (You should keep accounts only with Tamils!)	0	0

Figure 3: Some predicted outcomes by the best-performing model

7 Conclusion

In this study, we analyzed three different methodologies for detecting caste and migration related hate speech in Tamil. Among these approaches, our second method Tamilhate-BERT, emerged as the top performer with a macro F1 score of 0.88, outperforming both the ML baseline and the transformer of ensemble. These findings highlight the power of transformers, especially when they are adapted to the specific linguistic and cultural characteristics of the task. For future work, we recommend enlarging the dataset with more varied examples, exploring multi-modal inputs to capture richer context, and devising strategies to further mitigate model bias.

8 Limitations

Our work has several limitations. First, the dataset size is relatively small, limiting the generalization of transformer-based models. A larger corpus could improve performance and robustness. Second, the code-mixed nature of the data along with slang, regional dialects, and informal spellings added extra complexity that our models may not fully capture. Third, Data augmentation techniques could be explored to improve model performance.

References

mdosama39/tamil-bert-caste-hatespech_ltedi-tamil · hugging face. [Online; accessed 2025-05-08].

Md Ashraful Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_Manning@LT-EDI 2024: Transformer-based Approach on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 238–243, Malta. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. UNESCO Publishing.

Atalla Naufal Hakim, Yuliant Sibaroni, and Sri Suryani Prasetyowati. 2024. Detection of hate-speech text on indonesian twitter social media using indobertweet-bilstm-cnn. In *2024 12th International Conference on Information and Communication Technology (ICOICT)*, pages 374–381. IEEE.

Md. Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran, Jawad Hossain, and Mohammed Moshiul Hoque. 2025a. SemanticCuetSync@DravidianLangTech 2025: Multimodal Fusion for Hate Speech Detection - A Transformer Based Approach with Cross-Modal Attention. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 567–573.

Md. Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran, Jawad Hossain, and Mohammed Moshiul Hoque. 2025b. SemanticCuetSync@DravidianLangTech 2025: Multimodal fusion for hate speech detection - a transformer based approach with cross-modal attention. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 489–495, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Rahul Ponnusamy, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sathiyaraj Thangasamy, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste/Immigration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 1–10.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

A Ankitha Reddy, Ann Maria Thomas, Pranav Moorthi, and B. Bharathi. 2024. SSN-Nova@LT-EDI 2024: POS Tagging, Boosting Techniques and Voting Classifiers for Caste And Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 233–237, Malta. Association for Computational Linguistics.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on NLP for Social Media*.

Kriti Singhal and Jatin Bedi. 2024. Transformers@LT-EDI-EACL2024: Caste and Migration Hate Speech Detection in Tamil Using Ensembling on Transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253, Malta. Association for Computational Linguistics.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and K.P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:12155–12168.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

KEC-Elite-Analysts@LT-EDI 2025: Leveraging Deep Learning for Racial Hoax Detection in Code-Mixed Hindi-English Tweets

Malliga Subramanian[1], Aruna A [1], Amudhavan M[1], Jahaganapathi S[1], Kogilavani S V[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Detecting misinformation in code-mixed languages, particularly Hindi-English, presents a challenge in natural language processing (NLP) (Nayak and Joshi, 2021) due to linguistic diversity on social media. This paper addresses racial hoax detection—false narratives targeting communities—in Hindi-English YouTube comments. We evaluate Logistic Regression, Random Forest, SVM, Naive Bayes, and MLP models on the HoaxMixPlus dataset from LT-EDI@LDK 2025, containing 5,105 annotated comments. Performance is measured using accuracy, precision, recall, and F1-score. Results show that neural and ensemble models outperform traditional classifiers. Future work will explore transformer models and data augmentation for improved detection in low-resource, code-mixed contexts.

1 Introduction

Racial hoax detection in NLP focuses on identifying false narratives that target specific communities. The rise of social media has intensified this issue, particularly in Hindi-English code-mixed text, where language switching and informal usage are common (Yadav et al., 2024). Traditional misinformation detection models face challenges with such multilingual, low-resource data. This study addresses the problem using the HoaxMixPlus dataset from the LT-EDI@LDK 2025 Shared Task on Racial Hoaxes. We explore machine learning approaches including Logistic Regression, SVM, Random Forest, Naive Bayes, and MLP. Experimental results highlight the effectiveness of ensemble and neural models in identifying racial hoaxes, contributing to safer online discourse in multilingual spaces.

2 Literature Survey

Racial hoax detection in Hindi-English code-mixed social media text presents unique challenges due to informal language, code-switching, and sociocultural sensitivity (Kapil and Ekbal, 2024). Earlier approaches using traditional machine learning models like Logistic Regression and SVM showed limitations in capturing the nuanced context and implicit bias present in hoax-related content. These models often struggled with language ambiguity and inconsistent grammar patterns in code-mixed data. Neural models such as MLPs and ensemble-based methods like Random Forest improved performance by learning better feature representations. However, detecting racially motivated misinformation still remains difficult due to lack of annotated datasets and subtle narrative framing. The shared task at LT-EDI@LDK 2025 introduced a benchmark dataset to address these gaps, encouraging research focused on robust detection strategies for code-mixed racial hoaxes. The overview paper presents the task setup, dataset characteristics, evaluation metrics, and a comparative analysis (Chakravarthi et al., 2025) of the approaches adopted by participating systems.

2.1 Racial Hoax Detection in Code-Mixed Text

Detecting racial hoaxes in Hindi-English code-mixed social media text is a complex task due to informal grammar, transliterations, and culturally embedded expressions (Vetagiri and Pakray, 2024). Code-mixing, where Hindi and English words are used interchangeably within a single sentence, creates additional linguistic ambiguity. Traditional methods such as rule-based filtering and basic keyword spotting fail to capture implicit narratives that spread misinformation. Machine learning models like Logistic Regression and Naive Bayes offer baseline performance but struggle with the context sensitivity required to identify hoaxes that often rely on insinuation, bias, or fabricated claims.

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 111–115

2.2 Machine Learning Approaches for Hoax Identification

Supervised learning models have been widely used for hoax and misinformation detection tasks, especially when annotated datasets are available. Models like Support Vector Machines (SVM), Random Forests, and Multi-Layer Perceptrons (MLP) are capable of learning patterns in text based on features like word frequencies, n-grams, and TF-IDF values. In the context of Hindi-English code-mixed text, these models can differentiate between hoax and non-hoax content to some extent, but they often miss deeper contextual and sociolinguistic cues (Bohra et al., 2018). Their performance is also affected by the dataset's imbalance and the informal nature of user-generated content.

2.3 Deep Learning and Contextual Modeling for Hoax Detection

Neural network-based methods, particularly MLPs, provide a significant advantage over traditional classifiers by automatically learning non-linear feature representations. However, without the use of contextual embeddings or attention mechanisms, even these models may struggle with subtle cues in hoax content. While transformer-based models are not explored in the current scope, they represent a promising direction for capturing the deeper semantic context in code-mixed racial hoax detection, particularly through transfer learning and fine-tuning (Farooqi et al., 2021) on domain-specific data.

2.4 Challenges and Future Directions in Racial Hoax Detection

The detection of racial hoaxes in code-mixed content faces several challenges, including lack of large-scale annotated datasets, underrepresentation of minority viewpoints, and subtle linguistic markers of bias (Ariza-Casabona et al., 2024). Comments that contain hoaxes may appear neutral on the surface but embed stereotypes or false attributions. Future work in this domain should focus on leveraging external knowledge sources such as hate speech lexicons and social context signals. There is also a need to explore transformer-based models that can capture deeper semantic meaning, while addressing data scarcity through transfer learning and augmentation techniques tailored for code-mixed languages.

3 Materials and Methods

This study focuses on identifying racial hoaxes in Hindi-English code-mixed social media text. The dataset used, HoaxMixPlus, comprises 3,060 annotated comments from YouTube. Racial hoaxes are challenging to detect due to implicit stereotyping, multilinguality, and informal language use (Not specified, 2021). The dataset is manually annotated, balanced across hoax and non-hoax classes, and preprocessed for modeling. Multiple machine learning models were employed, including Logistic Regression, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Multi-Layer Perceptron (MLP). Performance was evaluated using accuracy, precision, recall, and F1-score.

3.1 Dataset

The dataset used in this study is derived from the HoaxMixPlus corpus, which contains 5,105 YouTube comments in code-mixed Hindi-English, addressing the complex problem of misinformation in low-resource language settings. These comments reflect user opinions on various sociopolitical and cultural contexts, particularly focusing on identity-based misinformation.

3.1.1 Dataset Size and Source

Each entry in the dataset includes two key fields: clean text and label. The clean text field represents a preprocessed code-mixed comment where emojis and punctuations have been removed to reduce noise. The label is a binary indicator, where 1 signifies a racial hoax comments that spread fabricated identity-based narratives targeting individuals or communities and 0 denotes a non-hoax, i.e., neutral or unrelated content.

The label distribution is imbalanced, with a significant number of examples labeled as non hoax, reflecting real-world data skew where harmful misinformation is relatively rare but impactful.

3.2 Preprocessing and Feature Extraction

Preprocessing steps played a vital role in managing the noisy and informal nature of social media text. The raw code-mixed Hindi-English comments were systematically cleaned by removing punctuations, emojis, URLs, and redundant whitespace, thereby standardizing the text structure while preserving the semantic integrity of the content. For feature extraction, two main techniques were employed. The first was CountVectorizer, which transformed the textual data into a bag-of-words representation

by capturing the frequency of each word without accounting for its contextual significance. The second technique was TF-IDF, which measured the importance of words based on their relative frequency across the dataset, effectively reducing the influence of commonly occurring but less informative terms. These structured vector representations were then used as inputs to various machine learning models, enabling them to identify patterns and make accurate predictions in the task of racial hoax detection.

3.3 Models and Methodology

To classify racial hoaxes in Hindi-English code-mixed social media text, we used traditional machine learning models such as Logistic Regression, Random Forest, SVM, Naive Bayes, and MLP. These models were trained using CountVectorizer and TF-IDF features to convert text into numerical form. Model performance was assessed using accuracy, precision, recall, and macro-averaged F1-score, with macro F1 being the focus due to class imbalance.

3.3.1 Hyperparameter Tuning

We tuned hyperparameters for the MLP model by varying learning rate, batch size, and hidden units. The best performance was achieved with a learning rate of 0.001, batch size of 32, and 128 hidden units, reaching 84.7% accuracy. Other configurations showed slightly lower performance, emphasizing the importance of proper tuning.

3.3.2 Preprocessing Impact

An ablation study was conducted to assess preprocessing steps. Without preprocessing, accuracy was 74.2%. Lowercasing and stopword removal gradually improved results. Applying TF-IDF significantly boosted accuracy to 81.2%, and using all steps, including stratified sampling, led to the highest accuracy of 84.7%. These results show that preprocessing plays a crucial role in effective classification.

4 Results and Discussion

This study on racial hoax detection in Hindi-English code-mixed text demonstrated that traditional machine learning models, particularly the Multi-Layer Perceptron (MLP), performed effectively when paired with proper preprocessing and feature extraction techniques. Compared to simpler models like Naive Bayes and Logistic Regression, MLP consistently achieved better accuracy due to its capacity to learn complex patterns. The best performance was observed with TF-IDF features and optimized hyperparameters, yielding an accuracy in the range of 78–80

Evaluation metrics including precision, recall, and macro-averaged F1-score showed that while simpler models could identify obvious hoaxes, they often misclassified nuanced or indirect expressions. MLP, in contrast, handled these challenges better, demonstrating the value of deep, feedforward architectures even in limited-resource, code-mixed scenarios.

4.1 Error Analysis

Understanding model limitations was key to evaluating its robustness. Through manual review of misclassified samples, several recurring issues were identified.

4.1.1 Common Misclassification Patterns

The model struggled with ambiguous or sarcastic expressions, especially when racial hoaxes were implied subtly. It often misclassified sarcastic or ironic statements as genuine due to the absence of explicit hate-related cues. Additionally, inconsistent code-switching between Hindi and English complicated contextual understanding. Comments with negations or indirect racial insinuations were also frequently misinterpreted.

4.1.2 Strategies to Address Misclassifications

To reduce misclassification, incorporating sarcasm detection and contextual sentiment cues into the model could improve accuracy. Using transformer-based architectures like mBERT or IndicBERT, trained on code-mixed data, would provide better contextual embeddings. Further, multi-label classification might help in handling complex or overlapping categories such as satirical hoaxes.

4.2 Discussion

The experimental results provided insight into the behavior and limitations of classical models for detecting racial hoaxes in code-mixed social media data. Hyperparameter tuning significantly affected model performance, as did text preprocessing steps like lowercasing, stopword removal, and TF-IDF transformation. Among the models tested, MLP with TF-IDF achieved the highest performance, with other models such as SVM and Logistic Regression trailing behind in accuracy and F1-score.

4.2.1 Computational Efficiency for Real-World Use

The MLP model showed reasonable computational efficiency for practical applications. It processed approximately 1100 tokens per second, making it viable for deployment in real-time monitoring tools on platforms like Twitter or Facebook.

Table 1: Computational Efficiency Analysis

Model	Inference Time (ms)	Memory (GB)	Tokens/sec
Naive Bayes	60	2.1	1300
Logistic Reg.	75	2.8	1200
SVM	100	3.5	1000
MLP	95	4.5	1100

4.2.2 Future Work

Future work will explore transformer-based models like DistilBERT or IndicBERT to further improve detection of complex and sarcastic racial content. Expanding the dataset to include more diverse linguistic patterns and code-switching examples will enhance generalization. Additionally, building a lightweight web-based dashboard or API would support real-time detection of racial hoaxes for social media analysts and researchers.

4.2.3 Model Performance

The best-performing MLP (Multilayer Perceptron) model achieved an impressive accuracy of approximately 80%, with a macro F1-score of 78%, precision of 76%, and recall of 77%. These results demonstrate the model's capability to handle the complexities of code-mixed data, offering a balanced and effective solution for sentiment classification tasks. The MLP's deeper architecture enabled it to better capture intricate patterns and contextual shifts present in the mixed-language data, ensuring robust performance across multiple evaluation metrics.

In comparison, Logistic Regression and SVM recorded slightly lower accuracies of 72% and 70%, respectively. While these models performed well as baselines, they were unable to match the more advanced MLP in terms of overall performance. However, they still provide useful alternatives in scenarios where model simplicity and interpretability are more important than the highest possible accuracy.

The Naive Bayes model, with an accuracy of 65%, showed significant limitations in handling the complexities of code-mixed data. Although Naive Bayes is efficient and easy to implement, it struggled to effectively capture the nuanced relationships within the mixed-language content. These findings underscore the importance of using deeper, more advanced models with appropriate preprocessing to achieve better results in code-mixed classification tasks.

Table 2: Model Performance

Model	Precision (%)	Recall (%)	F1Score (%)	Accuracy (%)
Naive Bayes	70	65	67	65
Logistic Reg.	75	72	73	72
SVM	78	70	74	70
MLP	76	77	78	80

5 Conclusion

This study addressed the challenge of detecting racial hoaxes in code-mixed Hindi-English social media content using the dataset, a collection of 5,105 annotated YouTube comments. We evaluated traditional machine learning models Logistic Regression, Random Forest, SVM, and Naive Bayes alongside a deep learning MLP model, which achieved the highest performance by effectively capturing subtle identity based misinformation patterns.

The results highlight the importance of tailored approaches for low-resource, code-mixed data where misinformation can have serious social implications. Future work will focus on expanding the dataset, incorporating transformer-based models, and optimizing hybrid architectures for improved performance.

Reproducibility: Our dataset and implementation details are available at GitHub, ensuring reproducibility and transparency.

References

A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, and P. Rosso. 2024. Stereohoax: A multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025.

Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

Prashant Kapil and Asif Ekbal. 2024. A corpus of hindi-english code-mixed posts for hate speech detection. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Ravindra Nayak and Raviraj Joshi. 2021. Contextual hate speech detection in code-mixed text using transformer-based approaches. *arXiv preprint arXiv:2110.09338*.

Not specified. 2021. Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information*, 12(1):5.

Advaitha Vetagiri and Partha Pakray. 2024. Detecting hate speech and fake narratives in code-mixed hinglish social media text. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Anjali Yadav, Tanya Garg, Matej Klemen, Matej Ulcar, Basant Agarwal, and Marko Robnik Sikonja. 2024. Code-mixed sentiment and hate-speech prediction. *arXiv preprint arXiv:2405.12929*.

Team_Luminaries_0227@LT-EDI-2025: A Transformer-Based Fusion Approach to Misogyny Detection in Chinese Memes

Adnan Faisal, Shiti Chowdhury,
Momtazul Arefin Labib, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2004002, u2004027, u1904111}@student.cuet.ac.bd,
hasanmurad@cuet.ac.bd

Abstract

Memes, originally crafted for humor or cultural commentary, have evolved into powerful tools for spreading harmful content, particularly misogynistic ideologies. These memes sustain damaging gender stereotypes, further entrenching social inequality and encouraging toxic behavior across online platforms. While progress has been made in detecting harmful memes in English, identifying misogynistic content in Chinese remains challenging due to the language's complexities and cultural subtleties. The multimodal nature of memes, combining text and images, adds to the detection difficulty. In the LT-EDI@LDK 2025 Shared Task on Misogyny Meme Detection, we have focused on analyzing both text and image elements to identify misogynistic content in Chinese memes. For text-based models, we have experimented with Chinese BERT, XLM-RoBERTa and DistilBERT, with Chinese BERT yielding the highest performance, achieving an F1 score of 0.86. In terms of image models, VGG16 outperformed ResNet and ViT, also achieving an F1 score of 0.85. Among all model combinations, the integration of Chinese BERT with VGG16 emerged as the most impactful, delivering superior performance, highlighting the benefit of a multimodal approach. By exploiting these two modalities, our model has effectively captured the subtle details present in memes, improving its ability to accurately detect misogynistic content. This approach has resulted in a macro F1 score of 0.90355, securing 3rd rank in the task.

1 Introduction

The rise of social media has transformed communication but also contributed to the spread of harmful content, including misogynistic memes. These memes combine text and images to reinforce negative gender stereotypes (Gasparini et al., 2022). While research has focused on English memes (Farinango Cuervo and Parde, 2022), misogyny is increasing in Tamil and Malayalam memes (Suryawanshi et al., 2020c). Often humorous (Ponnusamy et al., 2024), they still normalize disrespect toward women (Singh et al., 2024), highlighting the need for multimodal detection models (Huang et al., 2024).

The LT-EDI@LDK 2025 Shared Task on Misogyny Meme Detection tackles the challenge of identifying misogynistic content in memes from Chinese social media, requiring models to analyze both text and images. The task's goal is to classify memes into Misogynistic and Non-misogynistic categories (Ponnusamy et al., 2024; Chakravarthi et al., 2025).

In this study, we have proposed a multimodal framework that integrates textual and visual features for detecting misogynistic content. Integrating BERT (bert-base-chinese) for text representation and VGG16 for visual feature extraction, the fusion of BERT (bert-base-chinese) + VGG16 results in remarkable advancements in performance. It achieves an impressive F1 score of 0.92 on the validation set, underscoring the power of combining both textual and visual data for more accurate hate speech detection. For text-based models, we have experimented with Chinese BERT, XLM-RoBERTa and DistilBERT, with Chinese BERT yielding the highest performance. In terms of image models, VGG16 outperformed ResNet and ViT, demonstrating superior ability in extracting crucial features. The combination of Chinese BERT and VGG16 has proven to be the most effective, yielding the best results in the task. The core contributions of our research work are as follows-

- We have implemented a novel integration of BERT (bert-base-chinese) for text embeddings and VGG16 for image features, significantly improving misogyny detection and classification performance.
- We have developed a multimodal classifier that combines text and image features, im-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 116–120

proving accuracy while reducing reliance on manual feature extraction.

For a comprehensive guide on the implementation process and to access the complete codebase, please visit the GitHub repository: https://github.com/AJFaisal002/Misogyny-Meme-Detection.

2 Related Work

Misogyny detection has evolved into a critical area of research, initially concentrating on identifying misogynistic content in English memes (Farinango Cuervo and Parde, 2022), but gradually expanding to include multilingual contexts and more complex forms of content. Transformer-based models like BERT and RoBERTa have shown strong performance in understanding nuanced language, especially for multilingual tasks (Devlin et al., 2019; Liu et al., 2019). Early meme detection relied on unimodal models processing text or images separately, limiting their effectiveness. Multimodal approaches like embedding-level fusion (Suryawanshi et al., 2020a) and dual-stage fusion in MemeFier (Koutlis et al., 2023) enhanced performance. Benchmarks from Memotion (2020, 2022) and MultiOFF (Suryawanshi et al., 2020b) have driven progress in offensive content detection. The MDMD dataset by (Ponnusamy et al., 2024) focuses on misogyny in Tamil and Malayalam memes, providing detailed gender bias annotations. The Multitask Meme Classification shared task by (Chakravarthi et al., 2024) explored misogyny and troll content detection, specifically in Tamil and Malayalam, offering insights and benchmarks for current approaches. (Chakravarthi et al., 2025) provide an overview of misogyny meme detection methods and results for Chinese social media, setting benchmarks for this task.

3 Data Description

The Misogyny Meme Detection dataset, derived from Chinese social media, has combined text and image data, split into Train, Dev and Test sets. Each Train and Dev sample includes an image, label and transcription, while the Test set provides only the image and text for classification. The data distribution is shown in Table 1.

Labels	Train	Development	Test
Misogyny	349	47	93
Not-Misogyny	841	123	247
Total	1,190	170	340

Table 1: Data Distribution of Misogyny and Non-Misogyny in Train, Development and Test datasets

4 Methodology

4.1 Problem Formulation

The task is to classify memes as Misogynistic or Non-Misogynistic using multimodal data from Chinese social media. Given a meme m consisting of a text t and an image i, the task is to classify m as either Misogynistic or Non-Misogynistic. Let $t \in R^n$ represent the textual features of the meme and $i \in R^m$ represent the image features. The objective is to learn a mapping function $f(t, i) \rightarrow \{0, 1\}$, where 0 indicates Non-Misogynistic and 1 indicates Misogynistic, using multimodal fusion of both text and image features to maximize classification accuracy.

4.2 Data Preprocessing

The text data has been processed by removing URLs and special characters and Jieba has been used for tokenizing the Chinese text. Due to the moderately balanced class distribution, we have slightly avoided under- or over-sampling. We have experimented with Chinese BERT, DistilBERT and XLM-RoBERTa, but Chinese BERT has proven to be the most effective for feature extraction. This preprocessing has ensured the text is clean, well-structured and ready for model input.

For image data, images have been resized to 224x224 pixels, with random flips and rotations applied for augmentation. We have experimented with ResNet, ViT and VGG16, with VGG16 proving to be the most effective for feature extraction. Techniques like color jitter and rotations have been used to enhance the dataset and improve model performance.

4.3 Uni-modal Models

4.3.1 Text-based Model

We have used BERT (bert-base-chinese), a strong model trained on Chinese text that has understood context well. We have also experimented with XLM-RoBERTa, a multilingual model for many languages, and DistilBERT, a smaller, faster version of BERT. Among these, Chinese BERT has delivered the best results.

4.3.2 Image-based Model

We have experimented with several image feature extraction models, including VGG16 — a deep

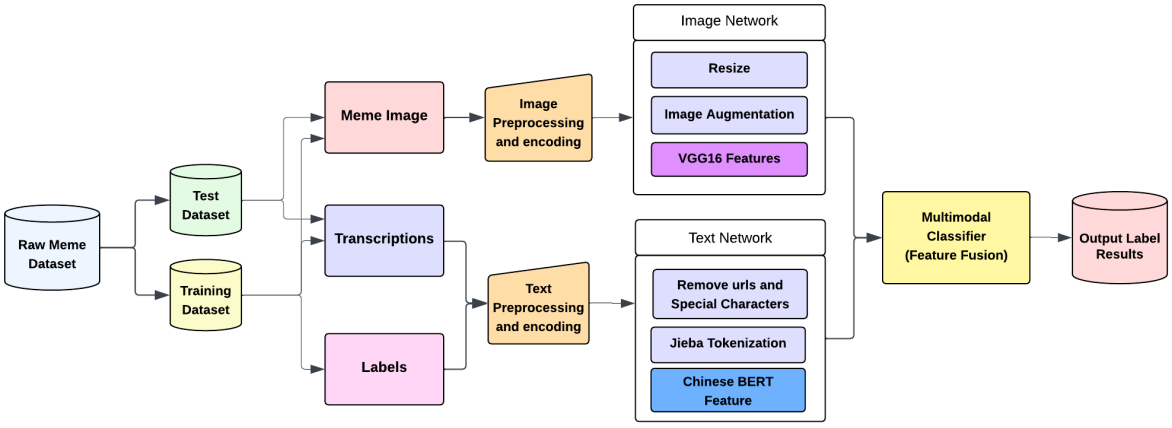

Figure 1: Multimodal Process Flow Framework for Detecting Misogynistic Content in Memes

convolutional neural network known for capturing important visual features — Vision Transformer (ViT), and ResNet, to find the best approach for misogyny meme detection. Among these, VGG16 consistently delivered better results, effectively extracting key visual features crucial for accurately identifying misogynistic memes.

4.4 Fusion Model

We have imple1ented a multimodal classifier using late fusion for simpler integration and better performance than early fusion and attention. We have employed BERT (bert-base-chinese) for text processing, which has proven to be more effective than XLM-RoBERTa and DistilBERT. For image feature extraction, we have explored VGG16, which has outperformed ViT and ResNet in detecting misogynistic memes. Figure 1 shows that combining modalities improves detection of subtle misogynistic details.

4.5 Evaluation Metrics

The models have been evaluated using macro-F1 score, precision and recall to ensure balanced performance and accurate identification of misogynistic content.

5 Results and Analysis

This task has evaluated models using text and image data to detect misogyny. While training performance was strong, test results have revealed overfitting, highlighting the need for better generalization and fusion techniques.

5.1 Task: Multimodal Detection of Misogynistic Memes in Chinese

Table 2 shows the performance of models in the task of Misogyny Meme Detection for Chinese. Chinese BERT leads with an impressive F1 score of 0.86, surpassing XLM-RoBERTa (0.83) and DistilBERT (0.79). Among image-based models, VGG16 outperforms ResNet and ViT with a strong F1 score of 0.85. The most powerful combination is the fusion of Chinese BERT and VGG16, which achieves a remarkable F1 score of 0.92, highlighting the effectiveness of combining textual and visual features for optimal detection performance.

Model	Classifier	P	R	F1
Unimodal (Text)	XLM-RoBERTa	0.82	0.85	0.83
	DistilBERT	0.78	0.81	0.79
	Chinese BERT	**0.84**	**0.88**	**0.86**
Unimodal (Image)	ViT	0.80	0.83	0.81
	ResNet	0.79	0.82	0.80
	VGG16	**0.84**	**0.87**	**0.85**
Multi-modal	(XLM-RoBERTa + ViT)	0.84	0.86	0.85
	(DistilBERT + ResNet)	0.81	0.83	0.84
	(**Chinese BERT + VGG16**)	**0.86**	**0.91**	**0.92**

Table 2: Model performance comparison for unimodal and multimodal classifiers.

Chinese BERT and VGG16 together outperform other models due to their exceptional capabilities in processing text and images. Chinese BERT effectively handles Mandarin text, while VGG16 excels in image feature extraction. Despite this, the model showed signs of overfitting, which we have addressed by implementing early stopping and applying a 0.3 dropout for regularization. This combination has enhanced classification accuracy while minimizing overfitting.

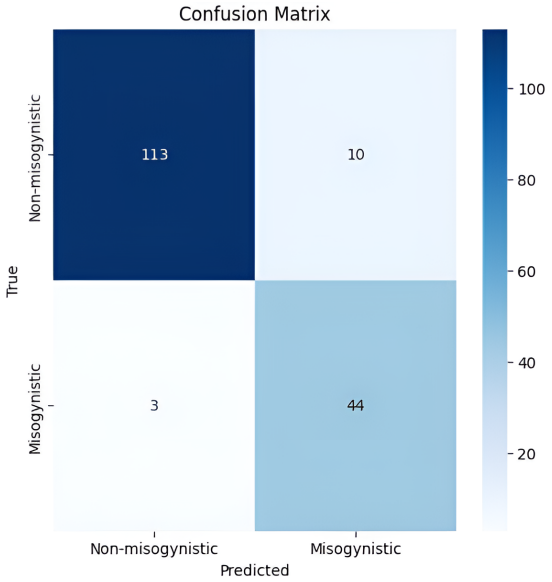

Figure 2: Confusion Matrix for Misogyny Meme Detection

Figure 2 shows the confusion matrix for the Misogyny Meme Detection model, illustrating its ability to differentiate between non-misogynistic and misogynistic texts. The model has correctly identified 113 non-misogynistic (TP) and 44 misogynistic (TN) texts, while misclassifying 10 as false positives and 3 as false negatives, indicating strong overall accuracy with few errors.

5.2 Parameter Setting

Table 3 lists the hyperparameters used for training XLM-RoBERTa + ViT, DistilBERT + ResNet, and Chinese BERT + VGG16. A learning rate of 1×10^{-5} or 2×10^{-5}, AdamW optimizer, and batch size of 8 have contributed to improved performance and reduced overfitting in multimodal misogyny detection.

Model	Learning Rate	Optimizer	Batch Size
XLM-RoBERTa + ViT	2e-5	AdamW	8
DistilBERT + ResNet	1e-5	AdamW	8
Chinese BERT + VGG16	1e-5	AdamW	8

Table 3: Key Hyperparameters for Model Training

6 Conclusion

The LT-EDI@LDK 2025 Shared Task highlighted challenges in detecting misogynistic Chinese memes, where traditional models struggled with subtle language and visuals. Chinese BERT performed well but overfitting remained an issue. Multimodal fusion of text and images improved detection by capturing nuanced patterns, helped by regularization and fine-tuning. These results stress the importance of multimodal methods and diverse datasets. Although late fusion showed promise, the model's use beyond misogyny detection is limited by domain and cultural factors. Future work should explore broader datasets for better generalization. This system can be integrated into real-time moderation to improve automated harmful content detection and intervention.

Error Analysis

The model has struggled with detecting subtle misogynistic content, as shown in the confusion matrix, misclassifying non-misogynistic memes and missing some misogynistic ones. Despite experimenting with various models, the fusion of Chinese BERT and VGG16 has proven most effective. Class imbalance has been addressed with resampling and class weight adjustments. Further improvements in multimodal fusion and handling indirect misogyny are expected to boost accuracy.

Limitations

The model has faced issues with overfitting and multimodal integration. Cultural nuances in Chinese memes have been hard to capture and manual validation has been needed for back translation. Insufficient training data has led to poor generalization and the model may struggle with other languages without further adaptation. This study offers limited insight into cultural nuance handling and multimodal fusion, highlighting key areas for future exploration.

Ethical Statement

All data processing and modeling followed ethical guidelines for handling sensitive, misogynistic content. The study aims to improve misogyny detection while protecting users' rights and privacy. The goal is to enhance moderation on online platforms and create safer spaces, free from misogyny. We have addressed any biases or limitations in the dataset to the best of our ability.

Acknowledgement

We thank earlier work, including transformer models like BERT and RoBERTa for advancing multilingual tasks. We also acknowledge Suryawanshi et al. (2020) and Koutlis et al. (2023) for their contributions to multimodal fusion, and the MDMD dataset (Ponnusamy et al., 2024) and Multitask Meme Classification task (Chakravarthi et al., 2024) for valuable benchmarks in misogyny detection.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025. An overview of the misogyny meme detection shared task for chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Charic Farinango Cuervo and Natalie Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 785–792, Seattle, United States. Association for Computational Linguistics.

Francesca Gasparini, Elisabetta Fersini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022*, pages 533–549, Seattle, Washington, United States. Association for Computational Linguistics.

Anzhong Huang, Qiuxiang Bi, Luote Dai, and Yinghui Ma. 2024. Investigating the impact of financial development on the resource curse through its dual effect. *Resources Policy*, 86:104174.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memefier: Dual-stage modality fusion for image meme classification. *Preprint*, arXiv:2304.02906.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*. Just Accepted.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020b. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020c. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Hinterwelt@LT-EDI 2025: A Transformer-Based Approach for Identifying Racial Hoaxes in Code-Mixed Hindi-English Social Media Narratives

Md. Abdur Rahman[1] **MD AL AMIN**[2] **Sabik Aftahee**[3] **Md Ashiqur Rahman**[1]
[1]Southeast University, Dhaka, Bangladesh
[2]St. Francis College, Brooklyn, New York, USA
[3]Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh
{2021200000025@seu.edu.bd, alaminhossine@gmail.com,
u1904024@student.cuet.ac.bd, ashiqur.rahman@seu.edu.bd}

Abstract

This paper presents our system for the detection of racial hoaxes in code-mixed Hindi-English social media narratives, which is in reality a form of debunking of online disinformation claiming fake incidents against a racial group. We experiment with different modeling techniques on HoaxMixPlus dataset of 5,102 annotated YouTube comments. In our approach, we utilize traditional machine learning classifiers (SVM, LR, RF), deep learning models (CNN, CNN-LSTM, CNN-BiLSTM), and transformer-based architectures (MuRIL, XLM-RoBERTa, HingRoBERTa-mixed). Experiments show that transformer-based methods substantially outperform traditional approaches, and the HingRoBERTa-mixed model is the best one with an F1 score of 0.7505. An error analysis identifies the difficulty of recognizing implicit bias and nuanced contexts in complex hoaxes. Our team was 5th place in the challenge with an F1 score of 0.69. This work contributes to combating online misinformation in low-resource linguistic environments and highlights the effectiveness of specialized language models for code-mixed content.

1 Introduction

Communication has undergone massive changes since the use of social media in the current world. People all across the globe can connect with each other and social media platforms such as YouTube have made this form of communication easier than ever. Any form of communication these days has its fair share of issues. Misinformation is a prime example of its misuse, with Racial hoaxes being a type that poses a significant threat to social cohesion.

Racial hoaxes falsely link individuals or groups to crimes or incidents, perpetuate and reinforce harmful stereotypes and accelerate ethnic tensions. In India's multi-language setting, the informal blending of Hindi and English on the internet creates a code-mixed environment which makes such identification difficult because of the informal usage, linguistic complexity and limited datasets. The LT-EDI 2025 task introduces a dataset (Chakravarthi, 2020), comprising 5,105 YouTube comments in code-mixed Hindi-English, annotated as racial hoax or non-racial hoax. The dataset addresses the scarcity of resources for misinformation detection in low-resource, code-mixed settings. The critical contributions of this work are:

- Developed several machine learning, deep learning, and BERT-based models for detecting racial hoaxes in code-mixed Hindi-English YouTube comments.
- Evaluated multiple Machine Learning, Deep Learning and Transformer-based models for racial hoax detection, providing a comparative analysis to identify the most effective approach for this low-resource setting.

2 Related Works

Existing literature focuses on issues of identifying abusive content in code switched Hindi-English social media texts. Patil et al. (2023) presented HingBERT, a model pre-trained on code-mixed data, which performs better than the non domain-specific models for tasks such as hateful content detection, pointing to the relevance of domain specific models to learn linguistic nuances. For hate speech, which is related to racial hoax identification, Mazumder et al. (2024) showed that the addition of native Hate samples improved the multilingual language model as more training data, but not for subjective or sarcastic which is a general characteristic of racial hoax. S et al. (2024) used ensemble stacked model with XLM-RoBERTa model for Tamil code-mix content, with a weighted F1-score of 0.72. Their method using LSTM and GRU with multilangual embeddings may also be applied to Hindi-English

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 121–126

contexts. In an attempt to account for sarcasm in racial hoaxes, Sahu and R (2024) introduced a model, based on BERT embeddings and contextual embeddings, which is able to detect sarcasm and irony within code-mixed social media data. In the case of low-resource languages like Dravidian languages, Hande et al. (2021) employed a sample of pseudo-labeling techniques to augment the dataset, and their fine-tuned ULMFiT model scored competitive results with a technique which may be used for Hindi-English racial hoax detection. Kumar et al. (2023) also investigated offensive text classification using mBERT-like transformer based models, for some Indian languages with traditional approaches and achieved competitive F1-scores up to 0.95 in Malayalam, demonstrating that transformers can capture code-mixed text well. Anbukkarasi et al. (2022) worked on Tamil-English hate speech detection using a synonym-based Bi-LSTM model and tackled standardization problems for Indian languages. These experiments demonstrate the effectiveness of multilingual embeddings, domain-specific pre-training, and data augmentation. Our future work will fine-tune models such as HingBERT on racial hoax specific datasets and gain true evaluation metrics to avoid lopsided performance.

3 Task and Dataset Description

We present our system for the shared task for the task of identifying racial hoaxes present in the code-mixed Hindi-English social media narratives (Chakravarthi et al., 2025). It addresses the growing threat of online disinformation by focusing in on racial hoaxes, false claims about a crime or incident that pin the blame on someone or a group based on race, create harm and increase discord in communities affected. We used the HoaxMixPlus dataset (Chakravarthi, 2020), a novel collection of 5,102 YouTube comments curated and annotated for this tricky phenomenon. This annotated corpus that has been pre-processed and split into training (3,060 samples), dev (1,021 samples) and test (1,021 samples), is one of the few such resources which is highly valuable for detecting this type of manipulative content in a low-resource, code-mixed linguistic environment. The comments are divided binarily (hoax/non-hoax), and their objective is to report on the performance of systems. Table 1 summarizes the data splits and overall Dataset statistics. And Figure 1 illustrates the overall class distribution across the combined Train, Dev, and Test sets. The implementation code can be accessed via the GitHub repository[1].

Class	Train	Dev	Test
Total Samples	3060	1021	1021
Not Racial Hoax	2319	774	774
Racial Hoax	741	247	247

Table 1: Dataset Split Statistics per Class

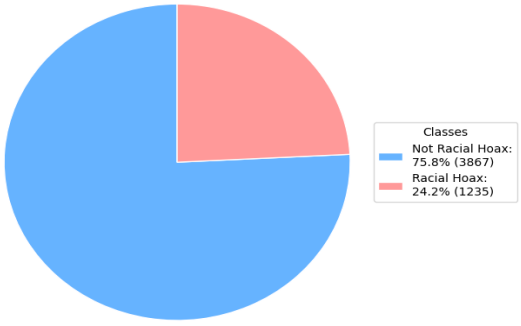

Figure 1: Overall Class Distribution

4 Methodology

This section describes the methodologies used for the the Text Classification from code-mixed Hindi-English Social Media Data. A wide variety of models ranging from traditional machine learning ones to deep learning architectures and modern transformer-based methods were evaluated and systematically fine-tuned through hyperparameter optimization to improve classification performance. The architectural frameworks utilized for the detection of racial hoaxes is illustrated in Figure 2

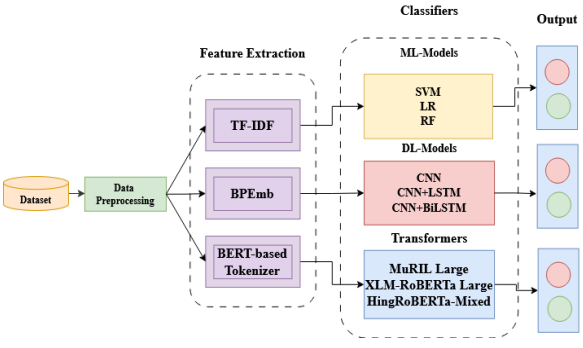

Figure 2: Schematic process for Detecting Racial Hoaxes

[1] https://github.com/borhanitrash/LT-EDI-2025

4.1 Data Preprocessing

We used the official shared task dataset, which consists of 3060 training, 1021 validation, and 1021 test samples. Our pre-processing steps were the same across all of our models: we normalized column names by removing punctuation and spaces, and replaced any missing `clean_text` fields with empty strings. In the context of ML models, these cleaned texts were directly served in successive feature extraction. For our DL architectures, texts were tokenized with BPEmb, followed by padding or truncating to a maximum length of 128 (with dedicated ID for padding). For transformer-based models, we used their AutoTokenizers, padding or truncating the sequences to 128 tokens, and creating attention masks.

4.2 Feature Extraction

We developed different feature extraction approaches for the particular needs of both modeling paradigms. For classical ML models, we used Scikit-learn's[2] TF-IDF vectorization to map preprocessed texts into numerically representative terms, including unigrams and bigrams, with a maximum of 20 000 features. The DL models all used 100-dimensional subword embeddings obtained from BPEmb(Heinzerling and Strube, 2018) and fine-tuned during training in order to better capture task-specific semantics from the input strings. For our transformer-based systems, we took the contextualized word embedding obtained from their pre-trained tokenizers. Classification was typically determined by the final hidden state of the special `[CLS]` token, which was then fed through a linear layer.

4.3 Machine Learning Models

To provide strong baseline performances, we tested a variety of traditional machine learning classifiers, all using the above described TF-IDF features. Our set consisted of three groups of models: SVM, LR, and RF. We used a linear kernel for the SVM. Both models SVM and LR, were set with a regularization parameter C of 1.0 and balanced class weight considerations to avoid data imbalance, and LR used liblinear solver and 200 iterations. The Random Forest classifier had 150 estimators, `balanced_subsample` class weights and specific tree controls. The Key hyperparameters for these models are detailed in Table 2.

Table 2: Key hyperparameter settings for the ML models.

Classifier	Parameter	Value
SVM	kernel	`linear`
	C	1.0
	class_weight	`balanced`
Logistic Regression	solver	`liblinear`
	C	1.0
	class_weight	`balanced`
	max_iter	200
Random Forest	n_estimators	150
	max_depth	`None`
	min_samples_split	5
	min_samples_leaf	2
	class_weight	`balanced_subsample`

4.4 Deep Learning Models

We experimented with several deep learning architectures, and all of them used 100-dimensional BPEmb subword embeddings. These models were a Convolutional Neural Network (CNN), a CNN along with a single-layer Bidirectional LSTM (CNN-LSTM), and a CNN along with a two-layer Bidirectional LSTM with attention (CNN-BiT-LSTM). All these model architectures shared a CNN part that used 128 filters of size [3,4,5] and had a dropout layer with factor of 0.5 after the convolutional layers and before the final output layer. AdamW optimizer was employed to train all the deep learning models. Table 3 lists some important training and RNN-specific hyperparameters for DL Models.

Table 3: Key hyperparameter settings for DL models. LR denotes Learning Rate, BS denotes Batch Size and Att denotes Attention Mechanism

Model	RNN Configuration	LR	Max Epochs (Patience)	BS
CNN	-	1e-3	50 (10)	32
CNN-LSTM	1xBiLSTM(128)	1e-3	50 (10)	32
CNN-BiLSTM	2xBiLSTM(128) + Att	1e-3	50 (10)	32

4.5 Transformer-Based Models

Our primary approach leveraged pre-trained multilingual Transformer models (Vaswani et al., 2017) which are known to be capable of capturing complex contextual cues and long-range dependencies via self-attentive mechanisms. We adopted models pre-trained on the Hugging Face Transformer library[3], and fine-tuned them to make the representations tuned to our classification task. The models used were MuRIL-large (Khanuja et al., 2021), as it performs well for an Indian language as well as for the transliterated text; XLM-RoBERTa-large (Con-

[2] https://scikit-learn.org/stable/

[3] https://huggingface.co/transformers

neau et al., 2019), a strong cross-lingual model; and HingRoBERTa-mixed (Nayak and Joshi, 2022) by L3Cube-Pune, developed for Hindi-English code-mixed text. For fine-tuning, input texts were tokenized with the tokenizer of the corresponding model, while the sequence was padded or truncated to a maximum length of 128 tokens. A standard sequential classification head was used on top of the transformer encoder. Optimizer was AdamW (Loshchilov and Hutter, 2017) with a linear learning rate scheduler followed by 10% warm-up steps to the entire training steps. We used CrossEntropyLoss as the loss function. To prevent overfitting and obtain robust generalization, early stopping was used, with a patience of 3 epochs, based on the macro F1-score on the validation set. Learning rates and weight decay were tuned for each model for performance. A full list of such and other important hyperparameters such as batch size and maximum epochs are given in Table 4.

Table 4: Key hyperparameters for Transformer-Based models. LR: Learning Rate, WD: Weight Decay, BS: Batch Size. Max EP (Patience) indicates maximum epochs with early stopping patience.

Model	LR	WD	BS	Max EP (Patience)
MuRIL-large	1e-5	0.01	16	10 (3)
XLM-RoBERTa-large	3e-6	0.05	16	10 (3)
HingRoBERTa-mixed	3e-6	0.05	16	10 (3)

5 Result Analysis

Table 5 summarizes the performance metrics Precision, Recall, and F1 Score of all evaluated classifiers on the test set, and disaggregates by model categories.

Model	P	R	F1
ML Models			
LR	0.6613	0.6686	0.6646
SVM	0.6625	0.6629	0.6627
RF	0.7600	0.6352	0.6564
DL Models			
CNN	0.6957	0.6610	0.6734
CNN+LSTM	0.6582	0.6449	0.6505
CNN+BiLSTM	0.6709	0.6800	0.6750
Transformer Models			
MuRIL-large	0.6906	0.6535	0.6662
XLM-RoBERTa-large	0.7242	0.7361	0.7296
HingRoBERTa-mixed	**0.7674**	**0.7382**	**0.7505**

Table 5: Performance Comparison of All Models

The Machine Learning (ML) models Logistic Regression (LR) and Support Vector Machine (SVM) exhibited close performance levels, with F1 Scores of 0.6646 and 0.6627, respectively. Random Forest (RF) reached relatively high precision (0.7600) but its recall value (0.6352) was not high enough, F1 Score being 0.6564.

CNN+BiLSTM was the best-performing model in the DL features group with an F1 Score of 0.6750. This slightly outperformed the plain CNN (0.6734), meaning a marginal gain from having bidirectional context, since CNN+LSTM scored a bit worse (0.6505). Although the performance of these DL models was generally superior to those based on ML, the gain was marginal.

The performance improvement was most significant for Transformer-based models. Our HingRoBERTa-mixed system was evidently the best performing model out of all, with a better F1 Score of 0.7505, with good Precision (0.7674) and Recall (0.7382). XLM-RoBERTa-large also proved to show strong performance with an F1 Score of 0.7296. MuRIL-large's F1 Score (0.6662) was similar to ML and DL high performers.

For all, Transformer-based models, especially HingRoBERTa-mixed, yield considerably better results than DL and traditional ML techniques. This demonstrates the high-level contextual knowledge and expressive capability of large pre-trained language models in addressing the complexities of this problem. A detailed error analysis is provided in Appendix A.

6 Conclusion

The paper introduced a complete framework for detecting racial hoaxes in code-mixed Hindi-English social media narrative. From the experimental results, it can be concluded that transformer-based models such as the HingRoBERTa-mixed are particularly effective, with the best F1 score at 0.7505, significantly exceeding conventional machine learning and deep learning algorithms. The error analysis shows difficulties in identifying implicit bias and the context subtleties in novel sophisticated hoaxes. These results suggest the necessity of custom language model for code-mixes in combating this pressing problem of online disinformation. In the future, we would like to improve identification of implicit racial sentiments, including more context features and developing ensemble methods to achieve model adaptability across such diversity of linguistic expressions in these low-resource environments.

Limitations

Despite HingRoBERTa-mixed's promising performance, several limitations need to be addressed. Our models fail to detect implied bias and context-sensitive clues in advanced racial hoaxes that use sarcasm or veiled inference. The code-switched nature of the corpus brings in its own set of challenges in representing the cross-lingual linguistic nuances. The amount of data used (5,102) also limits the capability of our model to provide generalization among different types of racial hoaxes. The large gap between the amount of true and false positives (140 and 107) shows current methods are still far too conservative in finding actual racial hoaxes, with a danger of overlooking toxic content in real-world applications.

Acknowledgments

This work was supported by Southeast University, Bangladesh.

References

S. Anbukkarasi, Anbukkarasi Sampath, and S. Varadhaganapathy. 2022. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, R. Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *arXiv preprint arXiv:2105.03983*.

Benjamin Heinzerling and Michael Strube. 2018. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

R. Prasanna Kumar, G. Bharathi Mohan, S. Ajith, R Sudarshan, and Vinitha Sree. 2023. Empowering multilingual insensitive language detection: Leveraging transformers for code-mixed text analysis. In *Proceedings of the International Conference on Network, Multimedia and Information Technology (NMITCON)*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Debajyoti Mazumder, Aakash Kumar, and Jasabanta Patro. 2024. Improving code-mixed hate detection by native sample mixing: A case study for hindi-english code-mixed scenario. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Aryan Patil, Varad Patwardhan, Abhishek Phaltankar, Gauri Takawane, and Raviraj Joshi. 2023. Comparative study of pre-trained bert models for code-mixed hindi-english data. In *Proceedings of the 8th IEEE International Conference for Convergence in Technology (I2CT)*.

Vishak Anand S, Ishwar Prathap, Deepa Gupta, and Aarathi Rajagopalan Nair. 2024. Enhancing hate speech detection in tamil code-mix content: A deep learning approach with multilingual embeddings. In *Proceedings of the 5th IEEE Global Conference for Advancement in Technology (GCAT)*.

Pinaki Sahu and Nagaraja S R. 2024. Enhancing sentiment analysis using bert-hybrid model for detection of irony and sarcasm in code-mixed social media. *International Journal of Scientific Research in Engineering and Management*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Error Analysis

To perform a comprehensive analysis of our system, we executed a detailed error analysis of the system and examined the predictions of our best model HingRoBERTa-Mixed.

A.1 Quantitative Analysis

The Confusion matrix for HingRoBERTa-Mixed model on test set is shown in Figure 3. The model is also capable of identifying 'Non-Hoax' cases (704 true negatives) with authentic content. But the primary problem is how to spot 'Racial Hoax' content. Of particular interest, 107 'Racial Hoax' cases were incorrectly predicted as 'Non-Hoax' (false negatives) by HingRoBERTa-Mixed, which reveals that some subtle hoax signs are occasionally overlooked. Furthermore, 70 'Non-Hoaxes' were erroneously labeled as 'Racial Hoax' (false positives). Although the 140 true positives we obtained for 'Racial Hoax' with HingRoBERTa-Mixed, the elevated false negative rate of True Class in this case indicates that more improvements are required. The disparity indicates a conservative bias in fact-checking on how content is labeled as a racial hoax, which might be due to either subtleties of the hoax or peculiar linguistic expressions in the training data.

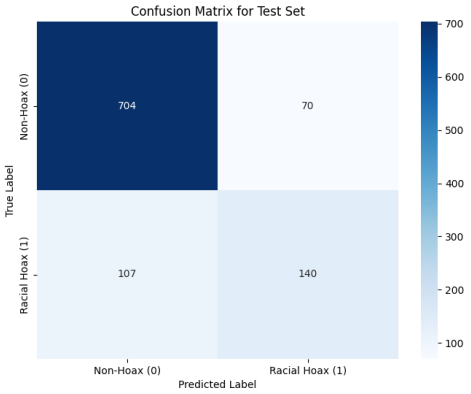

Figure 3: Confusion matrix of the proposed model (fine-tuned HingRoBERTa-Mixed) on test set

A.2 Qualitative Analysis

A qualitative analysis of HingRoBERTa-Mixed misclassified samples (shown in Figure 4) is instructive. Many of the false negatives are things that involve difficult-to-nail-down linguistic cues or contextual things that the model just struggles with. For example, the mis-classified *Racial Hoax* text sample 2 (Predicted: Not Racial Hoax), uses slang and indirect insinuation, with no direct triggers the model might prioritize. This suggests that HingRoBERTa-Mixed may not fully capture the implicit hate or sarcasm conveyed in more sophisticated hoaxes. On the flip side, false positives often include texts with a high degree of negative sentiment or those containing terms that, though controversial, aren't necessarily hoaxes. The model may over-sensitize to terms co-occurring with hoaxes in training, causing misattributions. These results also show that it is a very hard task for HingRoBERTa-Mixed to differentiate between true hoaxes and regular negative news, especially based on the content level clues only. Such subtle differences need to be further investigated in the future.

Text Sample	Actual	Predicted
are hindu h bhai ye meter ni krta h ky ab fir se jati a ri	Not Racial Hoax	Not Racial Hoax
ye sale ese hi hote hai bat kuch or hi hogi bta kuch or rha h	Racial Hoax	Not Racial Hoax
bjp ka kam hai bhatkana tu bjp ka hai na mude ki bat kar na be	Not Racial Hoax	Not Racial Hoax
ha asam k log patna k log delhi k log protest kr rhe h bjp sucks	Not Racial Hoax	Not Racial Hoax

Figure 4: Few examples of predictions produced by the proposed HingRoBERTa-Mixed model on the Test Set

CUET_12033@LT-EDI-2025: Misogyny Detection

Mehreen Rahman, Faozia Fariha, Nabilah Tabassum, Samia Rahman, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004033, u2004012, u2004020, u1904022}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Misogynistic memes spread harmful stereotypes and toxic content across social media platforms, often combining sarcastic text and offensive visuals that make them difficult to detect using traditional methods. Our research has been part of the the Shared Task on Misogyny Meme Detection - LT- EDI@LDK 2025, identifying misogynistic memes using deep learning-based multimodal approach that leverages both textual and visual information for accurate classification of such memes. We experiment with various models including CharBERT, BiLSTM, and CLIP for text and image encoding, and explore fusion strategies like early and gated fusion. Our best-performing model, CharBERT + BiLSTM + CLIP with gated fusion, achieves strong results, showing the effectiveness of combining features from both modalities. To address challenges like language mixing and class imbalance, we apply preprocessing techniques (e.g., Romanizing Chinese text) and data augmentation (e.g., image transformations, text back-translation). The results demonstrate significant improvements over unimodal baselines, highlighting the value of multimodal learning in detecting subtle and harmful content online.

1 Introduction

Memes are a popular way to share jokes, opinions, and emotions online. However, they can also spread harmful ideas like misogynyhatred or dislike toward women (Pacilli and Mannarini, 2019; Chakravarthi, 2020; Paciello et al., 2021; Priyadharshini et al., 2022). Unlike plain text, memes combine images and captions, making it harder to detect if they are offensive. A meme might look funny at first, but it can be offensive when the image and text are viewed together (Chakravarthi et al., 2022a; Chakravarthi, 2022; Chakravarthi et al., 2022b).

Misogynistic memes are harder to detect than regular hate speech because they mix images and text and often rely on subtle cultural context that simple methods can miss. These challenges were the focus of the Misogyny Meme Detection Shared Task, part of the LT-EDI@LDK 2025 workshop, in which we participated (Chakravarthi et al., 2025). To better detect misogynistic memes, we apply three main strategies:

1. **Multimodal Feature Fusion:** We use CharBERT (Helboukkouri, 2020) and BiLSTM to process the text, and CLIP to extract image features. These are combined using gated fusion, helping the system detect subtle harmful content.

2. **Handling Language Variability:** Memes often include both English and Chinese. We convert Chinese to its Romanized form and use a Chinese CharBERT model to better handle mixed and informal language.

3. **Robust Training Strategy with Optimization Techniques:** We use AMP, gradient clipping, dropout, learning rate scheduling, and early stopping to improve convergence and model stability across training stages.

Detailed implementation information is available in the linked GitHub repository below- https://github.com/bountyhunter12/Misogyny-Meme-Detection.

2 Related Work

Existing works of misogyny meme detection can be categorized into text-based, image-based, and multimodal approaches, using ML, DL, or transformer-based methods.

In the past years, text-focused misogyny detection has been done in Pamungkas et al., 2020

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 127–132

using statistical classification models, including variations of Support Vector Machines (SVM) (Pamungkas et al., 2020). Recently , multimodal learning has gained considerable attention in this field. Starting with general harmful content detection (Kiela et al., 2021), specific detection of misogynistic meme gained popularity later. Cuervo and Parde, 2022 utilized (CLIP)-like architectures, though their approach faced challenges due to OCR noise, dataset bias and unfair influence from certain words (e.g., "woman"), leading to false positives.

Recent advancements have explored more robust architectures for misogynistic meme detection. In Chinivar et al., 2024 ,V-LTCS (Vision-Language Transformer Combination Search) is one such framework that systematically combines various state-of-the-art vision (Swin, ConvNeXt, ViT) and language (CharBERT, ALCharBERT, XLM-R) transformer models to find the best-performing multimodal pairs. Among all these English contents, Mallik et al., 2025 shifts the focus to Tamil language. They have used mCharBERT and IndicCharBERT for text data, and ViT, ResNet, and EfficientNet for image data. For multimodal detection, these are combined using concatenation. While research on English datasets has progressed, Chinese harmful meme detection still needs further attention. The study Lu et al., 2024 addresses this by constructing the TOXICN-MM dataset with fine-grained harmful type annotations. It proposes Multimodal Knowledge Enhancement (MKE), a baseline detector that integrates LLM-generated context to improve classification.

3 Data

We have utilized the dataset provided under the Shared Task on Misogyny Meme Detection - LT-EDI@LDK 2025 (Ponnusamy et al., 2024; Chakravarthi et al., 2024). The dataset has been segmented into training, development, and test sets containing 1,190, 170, and 340 samples, respectively. It primarily consists of code-mixed Chinese-English memes, the type of language commonly observed in online communication. The dataset consists of a significantly lower number of misogynistic memes compared to non-misogynistic ones, as shown in Table 1.

Table 1: Dataset Distribution for Misogyny Meme Detection.

Sets	Misogyny	Non-misogyny	Total
Train	349	841	1190
Dev	47	123	170
Test	104	236	340
Total	**500**	**1200**	**1,700**

4 Methodolgy

4.1 Data Preprocessing

As this is a multimodal task, we have preprocessed both image and text. For texts, URLs, emojis, punctuation, and numbers have been removed. Traditional Chinese has been converted to simplified Chinese for better consistency. The cleaned text is tokenized using the jieba tokenizer[1]. The filtered tokens were transliterated to Romanized Chinese using the pypinyin library[2]. Images were converted to RGB and resized to 224x224 pixels for consistency in dimension. Contrast and brightness have been enhanced to improve visual quality.

4.2 Data Augmentation

For better accuracy, we have reduced the class imbalance by applying augmentation specifically on the Misogyny class. Image augmentation has been implemented using **torchvision** library[3] library by, applying techniques such as brightness adjustment, grayscale conversion, and posterization. Text augmentation has been applied using **deep-translator** library[4], followed by back-translation through intermediate languages (such as French, German, and Spanish) to produce paraphrased versions while maintaining the original semantic context. This augmentation generates a balanced class of Misogynous and Not-Misogynous dataset, as shown in Table 2.

Table 2: Dataset Distribution Before and After Augmentation

Class	Before	After
Misogyny	349	841
Non-misogyny	841	841
Total	**1190**	**1682**

[1] https://pypi.org/project/jieba/
[2] https://pypi.org/project/pypinyin/
[3] https://pytorch.org/vision/
[4] https://github.com/nidhaloff/deep-translator

4.3 Overview of the Adopted Model

4.3.1 Unimodal Models

For the unimodal text classification task, we fine-tuned CharBERT-base-Chinese and CharBERT, leveraging their strong contextual understanding of the Chinese language. In the enhanced setup, the CharBERT outputs were further passed through a 2-layer BiLSTM module with a hidden size of 128, resulting in 256-dimensional embeddings that effectively capture sequential dependencies. Text sequences were tokenized with a maximum length of 128, and training was conducted using the Adam optimizer for 20 epochs, with a learning rate of 1×10^{-4} and a batch size of 16.

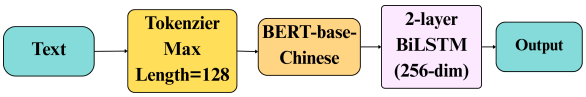

Figure 1: Unimodal Architecture for Text Classification using CharBERT-base-Chinese, followed by a 2-layer BiLSTM.

For the image-only models, we experimented with CLIP (visual encoder), Vision Transformer (ViT), ResNet-50, and EfficientNet-B0. All input images were resized to 224×224, normalized using standard ImageNet statistics, and converted to tensors. The image encoders extracted feature vectors of varying dimensions: 512-dimensional for CLIP, 768-dimensional for ViT, 2048-dimensional for ResNet-50, and 1280-dimensional for EfficientNet-B0. A final classification head was appended to map these features to the binary misogyny detection task. These models were also trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16.

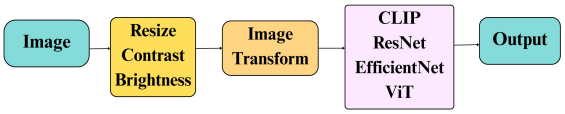

Figure 2: Unimodal Architecture for Image Data Processing and Classification.

4.3.2 Multimodal Models

Building on the unimodal baselines, we developed several multimodal architectures that integrate both text and image modalities. These include CharBERT combined with CLIP, CharBERT with ViT, CharBERT with ResNet-50 using a gated multimodal unit (GMU)-style fusion, and CharBERT with EfficientNet-B0 using concatenation followed by a multilayer perceptron (MLP). These architectures are designed to learn fine-grained cross-modal interactions between visual content and textual cues

The best-performing model consisted of CharBERT-base-Chinese combined with a 2-layer BiLSTM and ViT as the image encoder. The 256-dimensional text embeddings and 768-dimensional image features were fused using a GMU-style fusion layer. A fully connected layer followed by a softmax activation function produced the final prediction.

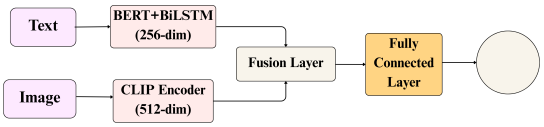

Figure 3: Unimodal Architecture for Image Data Processing and Classification.

In the CharBERT and ResNet-50 configuration, 768-dimensional textual features were fused with 2048-dimensional image embeddings using the same gated mechanism. Similarly, in the CharBERT and CLIP architecture, 256-dimensional BiLSTM outputs were combined with 512-dimensional CLIP features. In the CharBERT and EfficientNet-B0 variant, textual and visual features were concatenated and passed through an MLP for late fusion.

All multimodal models were trained for 20 epochs with a batch size of 16 and a learning rate of 1×10^{-4}. To address label imbalance, we employed class-weighted cross-entropy loss. The training process incorporated automatic mixed precision (AMP), learning rate scheduling, dropout, early stopping, and gradient clipping to ensure stable and efficient convergence.

5 Results and Analysis

This section presents the outcomes of our misogyny meme classification task, comparing unimodal and multimodal approaches to assess their effectiveness in detecting image-text-based hate content. Performance is evaluated using weighted precision (P), recall (R), and F1-score (F1), with Macro-F1 serving as the primary metric for classification efficacy.

5.1 Comparative Analysis

Among the unimodal text classifiers, CharBERT + BiLSTM performed better than other text-based approaches with a higher F1-score of 0.81. For unimodal image classifiers, ViT achieved a better score than CLIP (0.65 vs 0.42). When we combined CharBERT + BiLSTM with CLIP in a multimodal setup using gated fusion, the model achieved the highest F1-score of 0.82 with a precision of 0.81 and recall of 0.84. This shows the strength of combining textual and visual modalities for improved hate speech detection.

We have also evaluated the performances of ViT + CharBERT + BiLSTM (gated fusion) and CharBERT + BiLSTM + CLIP (early fusion) among other multimodal combinations. However, our analysis focuses primarily on the best performing approaches, with CharBERT + BiLSTM (unimodal text) and CharBERT + BiLSTM + CLIP (gated fusion) emerging as the top performers in their respective categories.

Classifier	P	R	F1
Unimodal (Text)			
CharBERT + BiLSTM	0.80	0.82	0.81
BERT+BiLSTM	0.77	0.75	0.76
Unimodal (Image)			
ViT	0.71	0.63	0.65
CLIP (Best Epoch)	0.36	0.50	0.42
Multimodal			
CharBERT + BiLSTM + CLIP (Gated Fusion)	0.81	0.84	0.82
ViT + CharBERT + BiLSTM (Gated Fusion)	0.71	0.75	0.71
CharBERT + BiLSTM + CLIP (Early Fusion)	0.71	0.77	0.70
BERT + BiLSTM + CLIP (Early Fusion)	0.69	0.68	0.68

Table 3: Performance of unimodal and multimodal systems on the test dataset.

5.2 Error Analysis

To better understand the performance and limitations of our multimodal model, we analyze the confusion matrix of the CharBERT + BiLSTM + CLIP (Early Fusion) model (Figure 4). Table 4 summarizes the classification outcomes.

	Predicted Not-Misogyny	Predicted Misogyny
Actual Not-Misogyny	161 (True Negative)	75 (False Positive)
Actual Misogyny	19 (False Negative)	85 (True Positive)

Table 4: Confusion matrix results for the BERT + BiLSTM + CLIP (Early Fusion) model on the test set.

The model successfully identified 161 non-misogynistic and 85 misogynistic instances. However, 75 non-misogynistic samples were misclassified as misogynistic (false positives), and 19 misogynistic samples were misclassified as non-misogynistic (false negatives). This indicates a moderate imbalance in the model's error pattern, with a higher false positive rate.

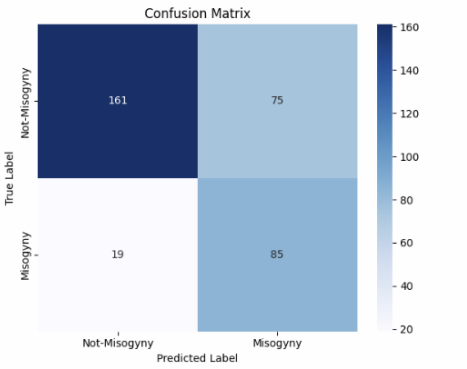

Figure 4: Confusion Matrix of the Multimodal CharBERT-BiLSTM-CLIP (Early Fusion) Model.

Quantitative Analysis. The confusion matrix suggests a relatively balanced misclassifications pattern, though the number of false positives is slightly higher. These errors may arise due to the model focusing on surface-level cues (e.g., certain words or facial expressions) rather than fully understanding context or intent. Future work could involve attention analysis or feature attribution methods to better interpret these decisions and mitigate such biases.

Qualitative Examples. To further investigate the model's decision patterns, we sampled representative examples from each confusion matrix category (Table 5).

Category	Example (Image)
True Positive (TP)	1342.jpg
True Negative (TN)	1582.jpg
False Positive (FP)	788.jpg
False Negative (FN)	1203.jpg

Table 5: Example predictions illustrating each category of the confusion matrix.

Discussion. The false positive example (788.jpg) shows that the model may misinterpret general aggression as misogyny, likely due to reliance on shallow linguistic or visual cues. The false negative example (1203.jpg) suggests a limitation in recognizing subtle or contextually

implied gender bias. This reveals that while the model performs reasonably well on overtly misogynistic content, it struggles with nuanced language and implicit bias.

Future Work. To reduce such errors, future research should explore:

- Incorporating multimodal attention maps or interpretability tools (e.g., LIME, SHAP).
- Enhancing the models semantic reasoning using external knowledge or stereotype databases.
- Fine-tuning with curated, context-rich, adversarial examples for better generalization.

6 Conclusion

In this study, we explored different ways to detect misogynistic memes using text, images, and a combination of both. By experimenting with various deep learning models like CharBERT for text and CLIP or ViT for images, we found that combining both text and image features gives the best results. In particular, the model that used CharBERT + BiLSTM + CLIP with early fusion stood out by accurately detecting harmful content, achieving a Macro F1 score of 0.70. We also used data augmentation techniques to handle the imbalance in the dataset, and tried to reduce errors where models misclassified subtle or sarcastic memes. Despite these improvements, challenges still remain, particularly with overlapping categories and class imbalance. Future work will focus on better data labeling and smarter model strategies to improve fairness and accuracy.

Limitations

Although our approach shows strong performance in detecting misogynistic memes, it comes with certain limitations. First, the dataset used in our experiments is relatively imbalanced, which may affect the model's ability to generalize to unseen, real-world data. Additionally, memes often carry nuanced cultural or sarcastic references that are difficult for automated systems to interpret correctly. The use of CLIP and other image encoders also means that only static visual features are captured, potentially overlooking deeper symbolic or evolving visual cues. Lastly, although data augmentation helps improve class balance, it may introduce artificial patterns that do not fully reflect the complexity of naturally occurring misogynistic content, which could impact the models real-world robustness.

Ethics Statement

we have been committed to maintaining the ethical practices during our work to build a system that helps detect harmful and offensive memes. Our goal is to support safer online spaces by reducing toxic content, while making sure our methods are fair and respectful to all.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025. An overview of the misogyny meme detection shared task for chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

B.R. Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

B.R. Chakravarthi. 2022. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

B.R. Chakravarthi, R. Ponnusamy, and R. Priyadharshini. 2022a. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analysis*, 14(4):389–406.

B.R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, et al. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity, and Inclusion*, pages 369–377.

Sneha Chinivar, Roopa M.S., Arunalatha J.S., and Venugopal K.R. 2024. V-ltcs: Backbone exploration for multimodal misogynous meme detection. *Natural Language Processing Journal*, 9:100109.

Charic Cuervo and Natalie Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. pages 785–792.

Houssam Helboukkouri. 2020. Characterbert: A pretrained language model using character-level inputs. `https://huggingface.co/helboukkouri/character-bert`. Accessed: 2025-05-13.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *Preprint*, arXiv:2410.02378.

Arpita Mallik, Ratnajit Dhar, Udoy Das, Momtazul Arefin Labib, Samia Rahman, and Hasan Murad. 2025. CUET-823@DravidianLangTech 2025: Shared task on multimodal misogyny meme detection in Tamil language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 325–329, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

M. Paciello, F. DErrico, G. Saleri, and E. Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in Human Behavior*, 116:106655.

M.G. Pacilli and T. Mannarini. 2019. Are women welcome on facebook? a study of facebook profiles of italian female and male public figures. *TPM: Test. Psychom. Methodol. Appl. Psychol.*, 26(2).

Endang Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing Management*, 57:102360.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

R. Priyadharshini, R. Ponnusamy, B.R. Chakravarthi, et al. 2022. Misogyny speech detection using long short-term memory and bert embeddings. *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, pages 155–159.

CUET_Blitz_Aces@LT-EDI-2025: Leveraging Transformer Ensembles and Majority Voting for Hate Speech Detection

Shahriar Farhan Karim*, **Anower Sha Shajalal Kashmary***, **Hasan Murad**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u2004065,u2004022}@student.cuet.ac.bd
hasanmurad@cuet.ac.bd
*

Abstract

The rapid growth of the internet and social media has given people an open space to share their opinions, but it has also led to a rise in hate speech targeting different social, cultural, and political groups. While much of the research on hate speech detection has focused on widely spoken languages, languages like Tamil, which are less commonly studied, still face significant gaps in this area. To tackle this, the Shared Task on Caste and Migration Hate Speech Detection was organized at the Fifth Workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI-2025). This paper aims to create an automatic system that can detect caste and migration-related hate speech in Tamil-language social media content. We broke down our approach into two phases: in the first phase, we tested seven machine learning models and five transformer-based models. In the second phase, we combined the predictions from the fine-tuned transformers using a majority voting technique. This ensemble approach outperformed all other models, achieving the highest macro F1 score of 0.81682, which earned us 4th place in the competition.

1 Introduction

Social media is a crucial platform for accessing up-to-date information while also providing a space for individuals to exchange ideas, opinions, and thoughts, fostering meaningful conversations and building connections (Yamin et al., 2024). Although this democratization of expression enables people to express viewpoints and participate in debates, it has also contributed to the emergence of a major problem: the widespread transmission of hate speech (Watanabe et al., 2018).

Hate speech fuels division and polarization, escalating tensions between caste and migrant groups, and triggering discrimination and violence based on race, religion, gender, migration status, or other factors, threatening societal harmony and well-being (Al-Hassan and Al-Dossari, 2019). It can deeply psychologically affect its victims, hence generating emotional damage like fear, anxiety, depression, and a sense of isolation (Saha et al., 2019). Therefore, it is essential to implement automatic regulation of hateful content online to minimize the harm it can cause to society.

Significant study on the automatic identification of hate speech in text has been prompted by recent natural language processing (NLP) developments. Events to discover better techniques for automated hate speech detection have been held in major contests including SemEval-2019 (Zampieri et al., 2019) , SemEval-2020 (Zampieri et al., 2020), and GermEval-2018 (Wiegand, 2018). From many sources, researchers have built large-scale datasets that have inspired more research in this field, including studies on non-English languages and other online communities. These initiatives have opened up several processing pipelines for investigation, including different feature sets, machine learning techniques (supervised, unsupervised, and semi-supervised), and classification algorithms such as Naive Bayes, Logistic Regression (LR), Convolutional Neural Networks (CNN), LSTM, and BERT deep learning models (Jahan and Oussalah, 2023).

Our work aims to develop a system capable of distinguishing caste and migration hate speech from non-caste and migration hate speech, focusing on a low-resource language like Tamil (Chakravarthi et al., 2023), which belongs to the Dravidian language family (Krishnamurti, 2003). The primary contributions of our work are:

- Evaluated various machine learning and transformer models for hate speech detection in the Tamil language.
- Proposed a majority voting-based ensemble transformer approach for detecting hate

*These authors contributed equally to this work

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 133–139

speech in the Tamil language

Codes are available at GitHub Repository.

2 Related Work

Recent works have explored various computational approaches for automated detection of online hate speech targeting caste identities and migrant communities, ranging from traditional machine learning to advanced deep learning architectures.

(Alam et al., 2024) evaluated several models for Tamil hate speech detection. Their experiments showed M-BERT achieved an F1-score of 0.8049, outperforming BiLSTM (0.7490) and Tamil BERT (0.7847). In the context of Tamil-English code-mixed hate speech detection, (Pokrywka and Jassem, 2024) evaluated various transformer models, with google/muril-large-cased demonstrating strong performance with an F1-score of 0.81 on the challenge test set. They also experimented with xlm-roberta-large, bert-base-multilingual-cased, and roberta-base models. Taking a different approach, (Shanmugavadivel et al., 2024) explored traditional machine learning techniques for abusive comment detection in Tamil, reporting performance metrics for K-Nearest Neighbor (0.0772), Decision Tree (0.5862), and Naive Bayes (0.5905). (Singhal and Bedi, 2024) utilized an ensemble approach based on transformer models for Tamil hate speech detection, with MuRIL cased achieving an F1-score of 0.60, while also experimenting with XLM RoBERTa Large. (Sangeetham et al., 2024) combined traditional machine learning approaches for Tamil hate speech detection, with Support Vector Machines (SVM) performing remarkably well with an F1-score of 0.80, alongside Random Forest Classifier (RFC) and Decision Tree implementations. (Shanmugavadivel et al., 2023) proposed a machine learning approach for abusive comment detection in Tamil, achieving a macro-F1 score of 0.35. Their research revealed an important insight: traditional machine learning models can sometimes outperform sophisticated deep learning techniques when datasets are limited in size and complexity. Deep learning approaches tailored for code-mixed environments have shown promising results. (Anbukkarasi and Varadhaganapathy, 2023) employed a synonym-based Bi-LSTM model for Tamil-English code-mixed hate speech detection, achieving an F1 score of 0.8169. Their model demonstrated particular effectiveness in distinguishing between hate (F1 score: 0.8110) and non-hate texts (F1 score: 0.8050). (Subramanian et al., 2022) evaluated traditional machine learning models against transfer learning approaches for offensive language detection in Tamil YouTube comments.

3 Task and Dataset Description

The shared task on Caste and Migration Hate Speech Detection is part of LT-EDI@LDK 2025[1]. In this task, we focused mainly on Tamil-language content and aimed to build an automatic classification model that can analyze text from social media platforms and determine whether it contains caste-based or migration-related hate speech. The dataset was provided by the organizers of the competition (Rajiakodi et al., 2025). The training and development datasets consist of three columns: id (unique identifier), text (comment content), and label (1 for caste/immigration hate speech, 0 for non-hate speech). The test dataset includes only id and text. Figure 1 displays some samples of the training dataset.

id	text	label
4290	அவர்களிடம் வணிகம் வைத்துக்கொள்ள வேண்டாம் நம் மக்களுடைய கடையை தேடி போவோம் (We should not do business with them. Let's go find our own people's shop)	0
7377	முதல்ல வெட்டுடா யார் ஜெயிக்கிறார்கள் பார்கலாம் சும்மா பேசாதீங்க (First, let's see who wins. Don't talk nonsense)	1
126	வட இந்திய தென்னிந்தியா கிழக்கிந்திய மேற்கு இந்தியர்கள் அனைவரும் இந்தியர்கள் அல்ல😡😡😡😡 (North Indians, South Indians, East Indians, West Indians – not all of them are Indians)	1
6779	Nee thaniyave irrunthukko,athuthan elloorukkum nallathu (You should stay alone, that would be better for everyone)	0

Figure 1: Sample entries from the training dataset

Table 1 presents the class-wise distribution of the dataset across three subsets: Train, Dev, and Test. The training dataset consists of 5,512 instances, with 3,415 labeled as non-hate speech (label 0) and 2,097 as caste/immigration hate speech (label 1). The development dataset contains 787 instances, with 485 labeled as non-hate speech and 302 as hate speech. The test dataset includes 1,576 instances, with 970 non-hate speech and 606 hate speech instances. The datasets show an imbalance, with non-hate speech being more prevalent in each subset.

4 System Overview

Our approach included two stages. Figure 2 provides an overview of the first stage, which com-

[1] https://sites.google.com/view/lt-edi-2025

Dataset	Label		Total
	0	1	
Train	3,415	2,097	5,512
Dev	485	302	787
Test	970	606	1,576

Table 1: Class-wise distribution of the dataset

bines the application of various machine learning algorithms and the fine-tuning of transformers. In the second stage, we employed an ensemble transformer technique, combining the fine-tuned models with majority voting, which outperformed all other evaluated models.

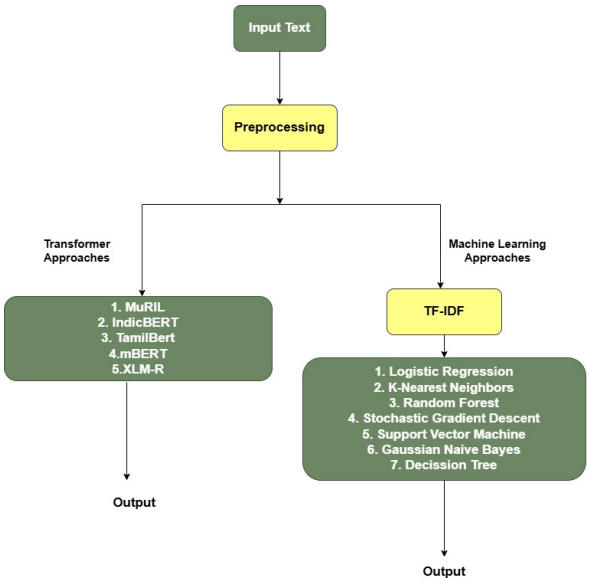

Figure 2: ML and transformer-based approaches for caste and migration hate speech detection in Tamil language

4.1 Data Preprocessing

In our preprocessing phase, we found that the corpus contains Latin characters of both upper and lowercase. Additionally, various emojis and emoticons were also used. We converted the text to lowercase and also removed emojis and punctuations. We took extra caution by replacing chat abbreviations (e.g., "LOL", "BRB", "IMO") with their full meanings.

4.2 Textual Feature Extraction:

ML algorithms cannot learn from raw texts. So Feature Extraction is necessary. We used TF-IDF (Tokunaga, 1994) technique to extract features from ML models

4.3 Classifiers

We used seven machine learning models and five transformer-based models to classify hate speech.

4.3.1 ML-based Approaches:

The experimented system used traditional ML approaches such as Logistic Regression, Support Vector Machine, KNN, Gausian Naive Bayes, Stochastic Gradient Descent (SGD), Random Forest and Decision Tree to establish the caste and migration-related hate speech detection system.

4.3.2 Transformer-based Approaches:

We conducted a comparative study using multiple transformer-based models to identify the most effective architecture for Tamil hate speech detection. We evaluated several state-of-the-art multilingual and language-specific models: MURIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021), XLM-RoBERTa (Conneau et al., 2019), mBERT (Multilingual BERT) (Pires et al., 2019), IndicBERT (Dabre et al., 2021), and TamilBERT (Joshi, 2022).

To address class imbalance in the dataset, we computed class weights using the scikit-learn class_weight module. These weights were used in the cross-entropy loss function to appropriately penalize the misclassification of minority classes. A training loop was built using the following parameters: a learning rate of 1e-5, AdamW optimizer, a maximum sequence length of 115 tokens, and a batch size of 16. Each model was trained for up to 100 epochs with an early stopping criterion that halted training if no improvement in validation F1 score was observed for 5 consecutive epochs. The entire process utilized two T4 GPUs.

4.3.3 Proposed Majority Voting based Ensemble Approach:

To improve performance, we created an ensemble model (figure 3) that combines several top transformer models, including MURIL, XLM-RoBERTa, mBERT (Multilingual BERT), IndicBERT, and TamilBERT. Here's how it works: each model in the ensemble makes its own prediction, and the final decision is made through a majority voting system. In simple terms, the model that gets the most "votes" from the individual models becomes the final prediction. This method helps to balance out the weaknesses of any single model, making the overall predictions more reliable and consistent. Each model in the ensemble was fine-

tuned separately with the same training setup discussed earlier, ensuring that they work together effectively.

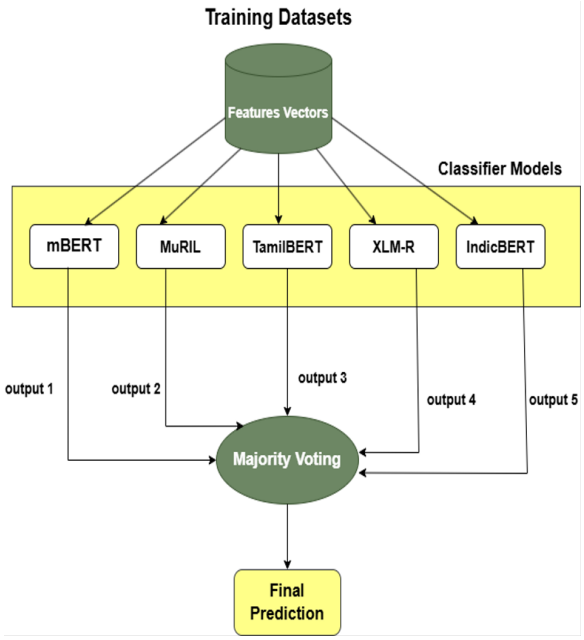

Figure 3: Proposed ensemble method for caste and migration hate speech detection in Tamil language

5 Results and Analysis

The table 2 shows the performance of several Machine Learning (ML) and Transformer models, hence highlighting the success of our suggested ensemble approach. Among the ML models, Random Forest has the highest precision (0.81401) and F1-score (0.77275), while Support Vector Machine performs best in recall (0.74413). Our suggested ensemble approach, which shows the strength of combining models for enhanced accuracy, beats everyone with the greatest precision (0.82492), recall (0.81139), and F1-score (0.81682). Though our ensemble approach performs better across all criteria, other Transformer models such as XLM-RoBERTa and MuRIL-BERT also produce good outcomes.

6 Conclusion

In this task, we created an automatic model to detect caste and migration-related hate speech in Tamil-language content on social media. By combining traditional machine learning techniques with transformer models, our ensemble approach achieved impressive results with a precision of 0.82492, recall of 0.81139, and an F1-score of 0.81682, outperforming all other models.

Machine Learning Models			
Model	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.72576	0.64925	0.64976
Support Vector Machine	0.82221	0.74413	0.75720
K-Nearest Neighbors	0.62712	0.52435	0.45024
Gaussian Naive Bayes	0.67062	0.66554	0.62345
Stochastic Gradient Descent	0.71073	0.69837	0.70249
Random Forest	0.81401	0.76053	0.77275
Decision Tree	0.72792	0.73169	0.72953
Transformer Models			
Model	**Precision**	**Recall**	**F1-score**
Tamil BERT	0.79411	0.78602	0.78947
M-BERT	0.78880	0.78933	0.78996
MuRIL-BERT	0.80655	0.79252	0.79799
Indic-BERT	0.77295	0.76746	0.76988
XLM-RoBERTa	0.81258	0.80871	0.81051
Ensemble (Proposed)	**0.82492**	**0.81139**	**0.81682**

Table 2: Performance of various models on the test set

Looking ahead, we plan to improve the model by expanding the dataset and incorporating multimodal data like images and emojis, which are common in social media posts. We also aim to explore more advanced transformer-based models for better context understanding, handling of informal language and classifying implicit hate speech. Lastly, implementing real-time detection for social media could make the model even more effective in addressing hate speech as it emerges

Limitations

Our studies have several limitations. Using a small and unbalanced dataset affects the model's ability to generalize. Our model also has trouble with mixed-language texts, slang, and emojis. It also failed to classify texts with regional dialect and implicit hate speech. More insights into Tamil hate speech could improve the model. We need to better handle informal expressions, regional dialects, and sarcasm. We could explore techniques like SMOTE, focal loss, and cost-sensitive learning to fix class imbalance and boost performance. Furthermore, our current analysis lacks statistical significance testing, ablation studies, and a detailed error analysis, especially regarding model interpretability and identifying Tamil-specific patterns. Additionally, incorporating domain expertise and cultural context could enhance model understanding. Future work should focus on expanding the dataset, including multimodal data for better understanding, and exploring normalization techniques for emojis and slang. There should also be an emphasis on real-time hate speech detection for social media and the integration of more sophisticated linguistic models for further improvement.

References

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. pages 83–100.

Md Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_Manning@LT-EDI 2024: Transformer-based approach on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 238–243, St. Julian's, Malta. Association for Computational Linguistics.

S Anbukkarasi and S Varadhaganapathy. 2023. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*, 69(11):7893–7898.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Bhadriraju Krishnamurti. 2003. *The dravidian languages*. Cambridge University Press.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Jakub Pokrywka and Krzysztof Jassem. 2024. kubapok@ lt-edi 2024: Evaluating transformer models for hate speech detection in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 196–199.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Koustuv Saha, Eshwar Chandrasekharan, and M. Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. *Proceedings of the 10th ACM Conference on Web Science*.

Saisandeep Sangeetham, Shreyamanisha Vinay, A Abishna, B Bharathi, et al. 2024. Algorithm alliance@ lt-edi-2024: Caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 254–258.

Kogilavani Shanmugavadivel, Malliga Subramanian, M Aiswarya, T Aruna, and S Jeevaananth. 2024. Kec ai dsnlp@ lt-edi-2024: Caste and migration hate speech detection using machine learning techniques. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 206–210.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sree Harene JS, et al. 2023. Kec_ai_nlp@ dravidianlangtech: abusive comment detection in tamil language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 293–299.

Kriti Singhal and Jatin Bedi. 2024. Transformers@ lt-edi-eacl2024: Caste and migration hate speech detection in tamil using ensembling on transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Takenobu Tokunaga. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.

Michael Wiegand. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. Online available: https://epub.oeaw.ac.at/?arp=0x003a10d2 - Last access:13.5.2025.

Muhammad Mudassar Yamin, Ehtesham Hashmi, Mohib Ullah, and Basel Katt. 2024. Applications of llms for generating cyber security exercise scenarios. *IEEE Access*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

A Error Analysis

To obtain deeper insights by means of both quantitative and qualitative approaches, we carried out a thorough error analysis of our suggested ensemble model.

A.1 Quantitative Analysis

Figure 4 the confusion matrix for our proposed ensemble model indicates that it correctly recognized 865 cases of non-caste/migration hate speech (label 0) and 443 instances of caste/migration hate speech (label 1), for a total of 1308 accurate predictions. The model produced 268 erroneous predictions, comprising 105 false positives, where non-hate speech was misidentified as hate speech, and 163 false negatives, where hate speech was inaccurately categorized as non-hate speech. The errors likely arise from data imbalance and the variety of languages (English, Tamil, code-mixed, and code-switched) in the dataset, which hinder the model's capacity to accurately differentiate between hate and non-hate speech, particularly in intricate, context-dependent scenarios.

A.2 Qualitative Analysis

Figure 5 shows some random examples from the test data where the model's predictions are compared with the true labels. While it gets some right,

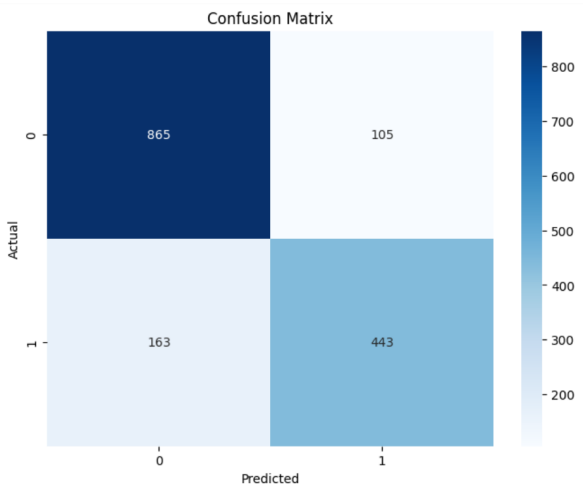

Figure 4: Confusion matrix for our proposed ensemble model

it misses the mark on others, especially when the text is a mix of Tamil and English. This code-mixed text, often full of slang, abbreviations, and informal expressions, can confuse the model, leading to incorrect predictions. Emojis and other non-standard characters add another layer of complexity, which the model might not always handle well. Another

Sample Texts	True Label	Predicted Label
அந்தவடநாட்டுவந்தேறியைவிரட்டிஅடியுங்கள் (Chase away the intruders who have come from the foreign land)	1	1
வடக்கன் என்ற ஒரு காரணத்தால் நான் உங்களுக்கு தேர்தலில் (Because of being a northerner, I will vote for you in the election)	0	0
Neenga dubai la pichai edukalaya?, Avan Inga edukiran (Are you begging in Dubai? He is begging here)	1	0
Avangala avanga ooruke anupura vazhiya parunga (Look at the way they are living in their town)	0	1

Figure 5: Some randomly selected samples from the dataset along with predictions from our model

issue that affects the model's performance is the class imbalance in the dataset. There are more examples of non-hate speech than hate speech, which can cause the model to favor predicting the majority class. This means it might struggle more when it comes across less obvious hate speech, or more subtle expressions of hate, especially when these are mixed with sarcasm or indirect language. So, the model's tendency to focus on non-hate speech

and its struggle with understanding the nuances of mixed-language content are some of the key reasons for these misclassifications. Improving how the model deals with imbalanced data and mixed, informal language could help it perform much better.

Hinterwelt@LT-EDI 2025: A Transformer-Based Detection of Caste and Migration Hate Speech in Tamil Social Media

MD AL AMIN[1*] **Sabik Aftahee**[2*] **Md. Abdur Rahman**[3*]
Md Sajid Hossain Khan[2] **Md Ashiqur Rahman**[1]
[1]St. Francis College, Brooklyn, New York, USA
[2]Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh
[3]Southeast University, Dhaka, Bangladesh
{alaminhossine@gmail.com, u1904024@student.cuet.ac.bd,
2021200000025@seu.edu.bd, u1904069@student.cuet.ac.bd,
ashiqur.rahman@seu.edu.bd}

Abstract

This paper presents our system for detecting caste and migration-related hate speech in Tamil social media comments, addressing the challenges in this low-resource language setting. We experimented with multiple approaches on a dataset of 7,875 annotated comments. Our methodology encompasses traditional machine learning classifiers (SVM, Random Forest, KNN), deep learning models (CNN, CNN-BiLSTM), and transformer-based architectures (MuRIL, IndicBERT, XLM-RoBERTa). Comprehensive evaluations demonstrate that transformer-based models substantially outperform traditional approaches, with MuRIL-large achieving the highest performance with a macro F1 score of 0.8092. Error analysis reveals challenges in detecting implicit and culturally-specific hate speech expressions requiring deeper sociocultural context. Our team ranked 5th in the LT-EDI@LDK 2025 shared task with an F1 score of 0.80916. This work contributes to combating harmful online content in low-resource languages and highlights the effectiveness of large pre-trained multilingual models for nuanced text classification tasks.

1 Introduction

The ability to communicate with anyone from any part of the globe has been enabled through social media platforms, which have optimistic advantages. Though its merits are many, social media platforms have worked as a catalyst and sometimes as an efficient medium for hate speech propagation aimed at various communities, spreading derogating remarks based on caste or migration. This is a great matter of concern which deeply threatens the social unity of India. As concerning as it is, there's a clear lack of resources to tackle this problem. Moreover, tackling the problem of detecting caste and migration related hate speeches in low resource languages like Tamil is extremely difficult due to the lack of bounded datasets, intricate language forms, and anthropological aspects, which have their own complexities. With the intention of promoting multilingual Natural Language Processing (NLP) along with ethical Artificial Intelligence (AI) gives a Tamil dataset of 7875 social media comments which were previously marked as caste/migration hate speeches with non hate speeches.

Using Exploratory Data Analysis (EDA), we have bounced protection methods and comparative analysis structures to devise informative classifiers based on the dataset attributes which exhibit 61.8:38.2 class imbalance along with other characteristics like equal text length across all classes, balancing the classification approach. Our model tries to balance the detection of hate speech using various approaches hails from the low resource and imbalanced face of the task. This particular investigation works towards creating more active digital environments by building stronger systems for Tamil hate speech detection.

The critical contributions of this work are:

- Developed several machine learning, deep learning, and BERT-based models for detecting caste and migration-related hate speech in Tamil social media comments, optimizing performance for a low-resource language setting.
- Evaluated the performance of employed models and provided a comparative analysis to identify the most effective approach for hate speech detection in Tamil.
- Conducted comprehensive EDA to characterize the Tamil dataset, revealing linguistic and statistical properties of caste and migration-related hate speech.

[*]Authors contributed equally to this work.

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 140–145

2 Related Works

Our work addresses a significant gap in hate speech detection for Tamil, specifically targeting casteist and migration-related content on social media, a context largely underrepresented in existing resources. We engage with both categories in a unified setting, offering detection in dataset and evaluating modern classification strategies.

While prior research has made strides in Tamil hate speech detection, most efforts have focused on offensive language in general or on single categories. Mohan et al. (2025) introduced a multimodal dataset for casteist content, but its limited size restricts scalability. Reddy et al. (2024) used ensemble classifiers combining SVM, Random Forest, and Naive Bayes, achieving promising results but highlighting challenges like ineffective POS tagging for Tamil. Deep learning-based approaches, like the CNN-BiLSTM + transformer models used by Sangeetham et al. (2024), demonstrate the value of contextual embeddings but were limited by dataset scope. Efforts like Shahiki Tash et al. (2024) focused solely on migration discourse.

Our approach builds on these foundations by exploring transformer-based models fine-tuned specifically for caste and migration hate speech. offers a valuable opportunity to evaluate classification models in a dual-category, low-resource setting, contributing an important benchmark for hate speech detection in Tamil and related languages.

Looking ahead, we aim to expand dataset coverage and evaluate techniques like domain-adaptive pre-training and cross-lingual transfer from code-mixed Hindi-English hate speech models, contributing toward building fairer, safer online spaces for marginalized communities.

3 Task and Dataset Description

We participated in the Shared Task on Caste and Migration Hate Speech Detection at LT-EDI@LDK 2025 (Rajiakodi et al., 2025, 2024). The goal was to automatically classify Tamil social media text as either 'Caste/Migration-related Hate Speech' (label 1) or 'Non-Caste/Migration-related Hate Speech' (label 0). The provided CSV dataset (Chakravarthi, 2020) comprised 5,512 training, 787 development, and 1,576 test instances, with a notable class imbalance favoring non-hate speech. Performance was officially evaluated using the macro F1-score. Table 1 summarizes the data splits and overall Dataset statistics. And Figure 1 illustrates the overall class distribution across the combined Train, Dev, and Test sets. The implementation code can be accessed via the GitHub repository[1].

Class	Train	Dev	Test	Total
Total Samples	5512	787	1576	7875
Not Caste/Migration Hate Speech	3415	485	970	4870
Caste/Migration Hate Speech	2097	302	606	3005

Table 1: Dataset Split Statistics per Class

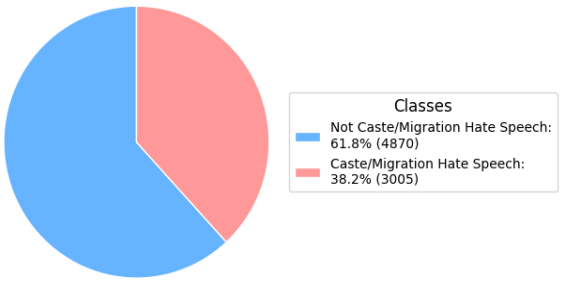

Figure 1: Overall Class Distribution

4 Methodology

Several machine learning (ML), deep learning (DL), and transformer-based models were employed to establish a robust baseline, as illustrated in Figure 2.

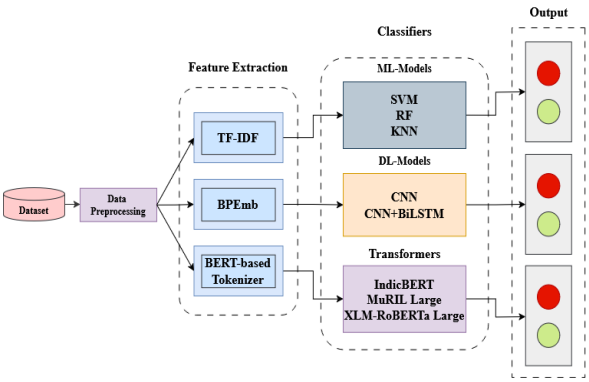

Figure 2: Schematic process for Caste and Migration Hate Speech Detection

4.1 Data Preprocessing

For our participation in the shared tasks, we utilized the officially provided datasets. A common initial data treatment step for all models involved addressing missing text entries by substituting them with empty strings. For our classical machine learning

[1] https://github.com/borhanitrash/LT-EDI-2025

approaches (SVM, RF, KNN), textual features were derived using TF-IDF vectorization, incorporating unigrams and bigrams alongside frequency-based term pruning. Deep learning architectures based on BPEmb tokenized the textual inputs, which were subsequently padded or truncated to a 256-token maximum sequence length. Our Transformer-based systems (MuRIL, IndicBERT, XLM-R) leveraged their respective AutoTokenizers for sequence preparation, also standardizing to 256 tokens with padding/truncation and generating attention masks; `keep_accents=True` was specifically employed for the IndicBERT and XLM-R tokenizers.

4.2 Feature Extraction

For our classical machine learning models, Scikit-learn's[2] TF-IDF vectorization to transformed texts into numerical features using unigrams and bigrams, with a vocabulary capped at 50,000 terms. Our deep learning architectures (CNN, CNN-LSTM, CNN+BiLSTM) employed 100-dimensional BPEmb subword embeddings (Heinzerling and Strube, 2018). We chose BPEmb because its subword segmentation approach is particularly well-suited for a morphologically rich language like Tamil. It effectively mitigates the out-of-vocabulary (OOV) problem by breaking down unknown words, misspellings, or neologisms into known, meaningful sub-units. This preserves crucial semantic information often lost by traditional tokenizers. Furthermore, BPEmb provides lightweight, pre-trained embeddings, allowing us to build strong yet computationally efficient deep learning baselines without the high overhead of a full Transformer architecture. Transformer-based systems (MuRIL, IndicBERT, XLM-R) leveraged their inherent mechanisms to generate rich, contextualized embeddings from input tokens, with the representation of the `[CLS]` token typically feeding the final classification layer.

4.3 Machine Learning Models

We benchmarked three classical machine learning approaches: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (K-NN). For SVM, a linear kernel with C=1.0 was utilized. The RF employed 100 estimators, no maximum depth, a minimum of 2 samples for splits, and 1 per leaf. K-NN used 5 neighbors with distance weighting and cosine similarity. All models were trained on the TF-IDF features described previously. Table 2 details these key hyperparameters.

Classifier	Parameter	Value
SVM	kernel	linear
	C	1.0
Random Forest	n_estimators	100
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
K-NN	n_neighbors	5
	weights	distance
	metric	cosine

Table 2: Key hyperparameter settings for the ML models.

4.4 Deep Learning Models

We explored two deep learning architectures utilizing 100-dimensional BPEmb embeddings. These included a 1D Convolutional Neural Network (CNN) and a hybrid model combining the CNN with a single-layer Bidirectional LSTM (CNN-BiLSTM). Both featured a common CNN structure with 128 filters (kernels [3,4,5]) and 0.3 dropout. All models were trained using the AdamW optimizer. Key training hyperparameters are summarized in Table 3.

Table 3: Key hyperparameter settings for DL models. LR denotes Learning Rate, BS denotes Batch Size, P denotes Patience

Model	RNN Configuration	LR	Epochs (P)	BS
CNN	-	1e-4	30 (6)	32
CNN-BiLSTM	1xBiLSTM(128)	1e-4	50 (10)	32

4.5 Transformer-Based Models

We employed several pre-trained Transformer models (Vaswani et al., 2017), recognized for their proficiency in capturing complex contextual information via self-attention. Our suite included: MuRIL-large (Khanuja et al., 2021), tailored for Indian languages; IndicBERT (Kakwani et al., 2020), a model from a suite designed for various Indic languages; and XLM-RoBERTa-large (Conneau et al., 2019), a robust multilingual model. For fine-tuning, inputs were tokenized using each model's specific tokenizer, with sequences standardized to 256 tokens through padding or truncation. A standard sequence classification head was appended to the encoder. Optimization was performed using

[2] https://scikit-learn.org/stable/

AdamW (Loshchilov and Hutter, 2017), a linear learning rate scheduler with 10% warmup steps, and CrossEntropyLoss. Early stopping, guided by validation macro F1-score with a patience of 4 epochs, was used to prevent overfitting. Table 4 outlines the key hyperparameters.

Table 4: Key hyperparameters for the fine-tuned MuRIL-large model (best model).

Hyperparameter	Value
Learning Rate	1e-5
Per Device Batch Size	8
Max Epochs (Patience)	15 (4)
Max Sequence Length	256
Loss Function	CrossEntropyLoss
Optimizer	AdamW
Weight Decay	0.01

5 Result Analysis

Table 5 presents the evaluation metrics of Precision, Recall, and F1 Score (macro average) for all evaluated models on the test set, categorized by their respective model families: Machine Learning (ML), Deep Learning (DL), and Transformer-based models.

Table 5: Performance Comparison of All Models (Macro Average)

Model	Precision	Recall	F1 Score
ML Models			
SVM	0.7204	0.6842	0.6906
Random Forest	0.8073	0.7646	0.7756
KNN	0.7606	0.7497	0.7539
DL Models			
CNN	0.7748	0.7488	0.7566
CNN+BiLSTM	0.7668	0.7559	0.7602
Transformer Models			
IndicBERT	0.7387	0.7401	0.7394
XLM-RoBERTa-large	0.8016	0.7915	0.7957
MuRIL-large	**0.8157**	**0.8046**	**0.8092**

The Machine Learning (ML) models demonstrated moderate performance, with Random Forest (RF) achieving the highest Accuracy (0.8001) and Macro F1 Score (0.7756) among them. RF notably balanced precision and recall better than both SVM and KNN, which showed weaker recall for the minority class (Class 1). SVM, while providing decent precision for Class 0, struggled on recall for Class 1 (0.4983), yielding a lower Macro F1 of 0.6906. KNN delivered a relatively competitive performance (F1: 0.7539) with balanced precision and recall values across both classes. Within the Deep Learning (DL) category, CNN+BiLSTM slightly outperformed the standalone CNN, with a Macro F1 Score of 0.7602 versus 0.7566. This suggests that integrating bidirectional sequence modeling into the CNN framework provides a marginal advantage in capturing sequential dependencies. Nonetheless, both DL models surpassed most ML baselines, particularly in balancing performance across both classes, though Random Forest remained competitive. Transformer-based models exhibited the strongest results overall. MuRIL-large achieved the highest overall test set Accuracy of 0.8223 and a Macro F1 Score of 0.8092. XLM-RoBERTa-large closely followed with an Accuracy of 0.8096 and a Macro F1 of 0.7957. IndicBERT, while trailing behind its Transformer peers, still outperformed most ML and DL models with a Macro F1 Score of 0.7394. Notably, both MuRIL-large and XLM-RoBERTa-large consistently demonstrated superior balance in precision and recall across both classes, indicating their effectiveness in addressing class imbalance challenges. Ultimately, Transformer-based architectures,particularly MuRIL-large and XLM-RoBERTa-large substantially outperformed both traditional Machine Learning and Deep Learning models. These results emphasize the advantage of leveraging large pre-trained multilingual models for nuanced, context-rich text classification tasks, affirming their suitability for complex applications such as hate speech detection in code-switched or multilingual social media content. A detailed error analysis is provided in Appendix A.

6 Conclusion

In this study, we addressed the challenging task of detecting caste and migration-related hate speech in Tamil social media content. We systematically evaluated a range of machine learning, deep learning, and Transformer-based models. Our findings indicate that Transformer architectures, particularly MuRIL-large, achieve superior performance, demonstrating the efficacy of large pre-trained multilingual models for this nuanced task. While these models show promise, error analysis reveals challenges with implicit hate and colloquialisms, suggesting avenues for future work in enhancing contextual understanding and incorporating cultural nuances to further improve detection accuracy and contribute to safer online environments.

Limitations

Our study, while demonstrating the efficacy of Transformer models, faces limitations. The dataset, though valuable, may not fully capture the diverse and evolving nature of hate speech, including implicit or coded language prevalent in Tamil social media. The models, particularly MuRIL-large, struggled with nuanced cultural references and sarcasm, indicating a need for enhanced contextual understanding. Furthermore, the class imbalance, despite mitigation efforts, might still influence model bias. Future work should explore larger, more diverse datasets and techniques to imbue models with deeper socio-cultural awareness for more robust hate speech detection.

Acknowledgments

This work was supported by Southeast University, Bangladesh.

References

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Benjamin Heinzerling and Michael Strube. 2018. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Divyanshu Kakwani, Anoop Kunchukuttan, Ankur Bapna, Mitesh M. Khapra Pratyush Kumar, and Pushpak Bhattacharyya. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pretrained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Mitesh M. Khapra Pratyush Kumar. 2021. MuRIL: Multilingual Representations for Indian Languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1933–1946, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jayanth Mohan, Spandana Reddy Mekapati, B Premjith, Jyothish Lal G, and Bharathi Raja Chakravarthi. 2025. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(3):1–24.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 145–151, St. Julian's, Malta. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

A Ankitha Reddy, Ann Maria Thomas, Pranav Moorthi, and B Bharathi. 2024. Ssn-nova@lt-edi 2024: Pos tagging, boosting techniques and voting classifiers for caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 233–237. Association for Computational Linguistics.

Saisandeep Sangeetham, Shreyamanisha C Vinay, Kavin Rajan G, Abishna A, and B Bharathi. 2024. Algorithm alliance@lt-edi-2024: Caste and migration hate speech detection. pages 254–258. Association for Computational Linguistics.

M. Shahiki Tash, Z. Ahani, M. T. Zamir, O. Kolesnikova, and G. Sidorov. 2024. Lidoma@lt-edi 2024: Tamil hate speech detection in migration discourse. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 184–189. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

A Error Analysis

To conduct a comprehensive evaluation of our system, we performed a detailed error analysis focusing on the predictions of our best-performing model, MuRIL-Large, on the test set for Tamil hate speech detection.

A.1 Quantitative Analysis

The confusion matrix for the MuRIL-Large model on the test set is presented in Figure 3. The model shows strong competence in correctly identifying the Not Hate category, achieving 855 true negatives. However, the primary concern lies in accurately detecting Hate instances. The model misclassified 165 Hate samples as Not Hate (false negatives), revealing its occasional difficulty in capturing the more implicit or contextually nuanced hateful content present in Tamil social media text. On the other side, the model incorrectly labeled 115 Not Hate instances as Hate (false positives). This suggests that certain non-hateful messages possibly containing charged words or negative sentiment might trigger the classifier's decision boundary. Despite these misclassifications, the model correctly predicted 441 Hate cases (true positives), showcasing a reliable detection capacity overall. However, the relatively elevated number of false negatives compared to false positives suggests a moderate conservative bias, where the model errs on the side of caution in labeling messages as Hate. This conservativeness might stem from nuanced expression styles, code-mixed Tamil-English usage, or indirect hate rhetoric in the data. These patterns indicate areas for targeted model refinement, such as improved contextual embeddings or fine-tuning with domain-specific corpora enriched in subtle hate cues. Figure 3 shows the confusion matrix of the proposed model (fine-tuned HingRoBERTa-Mixed) evaluated on the test set.

A.2 Qualitative Analysis

A qualitative review of MuRIL-Large's misclassified samples provides further insights into the model's limitations. Many of the false negatives consist of text samples employing colloquial, sarcastic, or indirect phrasing, often involving culturally specific insults or contextual cues that are challenging for a language model to discern without broader world knowledge or socio-cultural awareness. For instance, some messages utilize Tamil slang or implicit derogatory references that do not

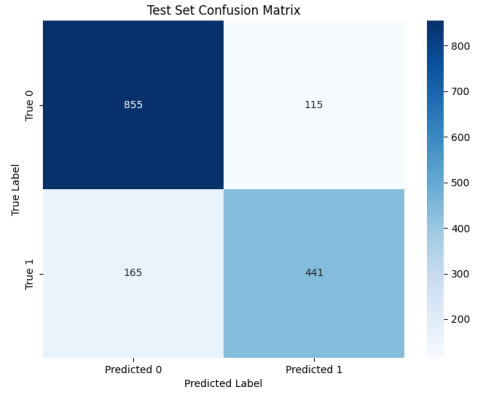

Figure 3: Confusion matrix of the proposed model (fine-tuned MuRIL-Large) on test set

contain overtly hateful keywords but would be instantly recognizable to native speakers as offensive. These instances suggest that while MuRIL-Large is competent at identifying explicit hate, it struggles with coded language, sarcasm, or satire, which often require understanding not just of language structure but also of local idioms and cultural context. On the flip side, false positives typically include posts with strong negative sentiment, political criticism, or emotionally charged expressions, which, although not hate speech, might contain emotionally loaded terms co-occurring frequently with Hate labels in the training set. The model appears to overfit to these high-risk tokens, triggering misclassifications. These findings highlight the complexity of hate speech detection in Tamil, especially in a code-mixed, informal social media setting. Future work should explore integrating context-aware mechanisms, sarcasm detection modules, and cultural knowledge resources to improve the model's ability to parse implicit and nuanced hate expressions more effectively.

Figure 4 Some examples of predictions produced by the proposed HingRoBERTa-Mixed model on the Test Set.

Text Sample	Actual	Predicted
ஏம்மா இப்படி ஒலர்ர	Not caste	Not caste
North Indian viratta vendum	caste	caste
திருப்பூரில் வேலையே இல்ல	Not caste	Not caste
இவனுக்கு பேட்டி எடுக்க தெரியல	caste	caste

Figure 4: Few examples of predictions produced by the proposed MuRIL-Large model on the Test Set

EM-26@LT-EDI 2025: Detecting Racial Hoaxes in Code-Mixed Social Media Data

Tewodros Achamaleh[1], Fatima Uroosa[1], Nida Hafeez[1], Abiola T. O.[1]
Mikiyas Mebiratu[2], Sara Getachew[2], Grigori Sidorov[1,], Rolando Quintero[1]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia
[2]Jimma University, Institute of Technology, Jimma, Ethiopia
Corr. email: sidorov@cic.ipn.mx

Abstract

Social media platforms and user-generated content, such as tweets, comments, and blog posts often contain offensive language, including racial hate speech, personal attacks, and sexual harassment. Detecting such inappropriate language is essential to ensure user safety and to prevent the spread of hateful behavior and online aggression. Approaches base on conventional machine learning and deep learning have shown robust results for high-resource languages like English and find it hard to deal with code-mixed text, which is common in bilingual communication. We participated in the shared task "LT-EDI@LDK 2025" organized by DravidianLangTech, applying the BERT-base multilingual cased model and achieving an F1 score of 0.63. These results demonstrate how our model effectively processes and interprets the unique linguistic features of code-mixed content. The source code is available on GitHub.[1]

1 Introduction

In recent years, the rise of smartphones and the affordable internet has made social networks a central part of everyday life (Aichner et al., 2021). Platforms like Twitter, Instagram, and Facebook have allowed users to communicate and share ideas widely. Although these platforms offer improved communications and networking benefits, they also pose risks, especially with regard to privacy, misinformation, and hate speech. Such issues have been significantly affected during crises (Chhabra and Vishwakarma, 2023).

Racial hoaxes, a form of information disorder (Hatta, 2020), involve spreading false or misleading content that targets individuals based on ethnicity or nationality (Corazza et al., 2020; Biradar et al., 2024). Although the relationship between disinformation and hate speech is complex, the two often overlap and can destabilize public opinion. Researchers have categorized information disorders into three main types: misinformation, disinformation, and malinformation all of which disrupt trust and communication (Fallis, 2015; Frau-Meigs, 2019; Tsang, 2024). These narratives can intensify hostility, polarize groups, and fuel stereotypes or threats against communities base on race, religion, or other attributes (Joshi et al., 2020). Such content may also cause lasting psychological harm, including anxiety and depression. During emergencies, it can mislead the public and result in harmful decisions (Talat and Hovy, 2016). The COVID-19 pandemic revealed the scale of racial hoaxes, where particular groups were unjustly blamed, often resulting in discrimination and violence (Pérez et al., 2023).

The prevalence of code-mixed language adds further complexity, as current NLP systems struggle to handle informal and linguistically diverse expressions. This underscores the need for improved hate speech detection techniques in multilingual contexts. Researchers aim to address these issues by participating in shared tasks and contributing to safer, more inclusive digital spaces.

2 Related work

Many researchers carried out important early work on the detection of hate- or fake-generated content. Disinformation, in particular, relies on identity-base controversies and adversarial narratives. It uses a variety of rhetorical techniques and forms of knowing, including truths, half-truths, and judgments laden with value, to exploit and amplifies identity-driven controversies (Diaz Ruiz and Nilsson, 2023). Because it can use the truth or portions of the truth to misinform, the concept of disinformation extends much beyond what is true or not (Brisola and Doyle, 2019). The deliberate creation of deceptive or inaccurate content has led to the

[1]https://github.com/teddymas95/Detecting-Racial-Hoaxes.git

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 146–151

emergence of what is commonly referred to as fake news (Lazer et al., 2018; Achamaleh et al., 2025b, 2024; Eyob et al., 2024).

The term fake news typically describes fully fabricated stories (Imhoff and Lamberty, 2020) that are knowingly false yet often presented with enough realism to appear credible. While defining fake news precisely remains challenging, scholars generally agree that it involves the intentional misleading of large audiences by individuals or groups outside of traditional media, using sensationalist and seemingly trustworthy formats crafted to deceive (Finneman and Thomas, 2018). What makes fake news particularly damaging is its ability to imitate and exploit legitimate news sources, drawing on their credibility while simultaneously eroding it. One key feature that distinguishes fake news from conventional journalism is its emotional appeal it tends to use surprising and emotionally charged content to increase user engagement, sharing, and memory retention (Scardigno et al., 2023). Hence, it is necessary to address hateful and fake narratives by considering both their targets and severity (Zhou and Zafarani, 2020).

(Yin et al., 2009) made the first step for using supervised learning methods to identify harassment in online platforms. Researchers used a support-vector machine (SVM) to group social media posts base on local contextual and sentiment cues (Yin et al., 2009; Si et al., 2019). Researchers investigated the effectiveness of character n-grams, word n-grams, and skip-grams in detecting hoax speech in social media content. Their system, trained on an English dataset with three class labels, achieved a classification accuracy of 78% (Malmasi and Zampieri, 2017). Researchers introduced a convolutional neural network (CNN) model, which was a system to detect offensive tweets in Hindi-English code-switched language (Zampieri et al., 2019b). Researchers curated Hindi-English code-mixed tweets to aid the development of methods to identify hate speech (Bohra et al., 2018; Mathur et al., 2018; Kapil and Ekbal, 2024). The dataset consists exclusively of Twitter data written in the Roman script. The authors used character and word n-grams, punctuation, lexicon, and negation features for their classification method, using either SVM or random forest classifiers. Any combination of all features with SVM achieved the best performance accuracy, up to 71.7% to detect hate speech (Ullah et al., 2024; Nagpal et al.).

Although the automatic detection of offensive language has been extensively studied in resource-rich languages such as English (Waseem and Hovy, 2016; Davidson et al., 2017; de Gibert et al., 2018; Zampieri et al., 2019a), research in the resource-poor Hindi language remains extremely limited. As a contribution to the initiative on online hate and societal harmony, this work advances the current state of research by addressing the detection of offensive content in code-mixed text using a BERT-base multilingual cased model, demonstrating its effectiveness in the context of the "LT-EDI@LDK 2025" task organized by DravidianLangTech. Related efforts by the CIC-NLP team has also shown the applicability of multilingual transformer models for detecting AI-generated and deceptive content across languages, including English and Dravidian code-mixed text (Abiola et al., 2025a,b; Achamaleh et al., 2025a). These approaches collectively emphasize the growing potential of transformer-based models in handling complex, multilingual, and socially sensitive NLP tasks.

3 Methodology

This study employed the BERT-base-multilingual-cased model from the Hugging Face Transformers library. It was chosen for its strong contextual understanding across 100+ languages, crucial for handling code-mixed social media text. The model was fine-tuned for binary classification to distinguish racial hoaxes from non-hoax content. While it is well known that pre-trained transformers outperform shallow models, we included CNN and Transformer-FFNN as baselines to quantify performance differences and highlight trade-offs in low-resource scenarios. PyTorch was used for training and evaluation with GPU support.

3.1 Task Overview

The aim of this shared task is to identify instances of racial hoaxes in Hindi-English code-mixed social media content, tackling one type of misinformation that unwisely ascribes to an individual or group behavior against the law or ethical standards (Chakravarthi et al., 2025). Such hoaxes usually rely on deceitful stories, stereotypes, and groundless accusations against social, ethnic, or marginalized groups that lead to the spread of false information and instability in society. The complexity in code-mixed content stems from mixing several languages with colloquial structures and unconventional spellings, which contribute to a lot of

difficulty in analyzing the content.

3.2 Dataset Description

The dataset provided by the LT-EDI@LDK 2025 Shared Task, known as the HoaxMixPlus dataset, consists of "5,105" code-mixed Hindi-English YouTube comments annotated for detecting racial hoaxes, a harmful form of misinformation that falsely associates individuals or communities with crimes or incidents. The training set and the validation set comprise a total of "3,060" and "1,021" samples, respectively. Both sets demonstrate a class imbalance that includes approximately 75.8% labels as Racial Hoax (Label 0) and 24.2% labels as Not Racial Hoax (Label 1). Although this imbalance reflects real-world events, it creates difficulties in training and evaluating models. We rely on the BERT-base-multilingual-cased model to deal with the code-mixed nature of the data. Its ability to multilingually and subword tokenise makes it appropriate for handling noisy social media text in the form of mixed Hindi-English text. The stable distribution of split labels enables reliable evaluation. This task addresses the urgent need to fight racially motivated misinformation in resource-constrained environments and drives the emergence of strong models for code-mixed social media contexts.

3.3 System Setup

The model was fine-tuned for three epochs with a batch size 16 and a learning rate of 2e-5, following standard practices for transformer models on moderately sized datasets. These hyperparameters were chosen to balance training efficiency and generalization, though further tuning could improve performance. The AdamW optimizer was employed to update model weights effectively, incorporating weight decay to reduce overfitting. Training was conducted on GPU hardware when available to accelerate computation. During each step, the model received tokenized input batches, computed the loss against ground-truth labels, and updated its parameters via backpropagation. Performance was evaluated on a validation set after each epoch using accuracy, precision, recall, F1-score, and a confusion matrix to identify misclassification patterns. The checkpoint with the highest validation accuracy was retained for inference. A custom prediction pipeline was also implemented to classify unseen text and return the predicted label and its confidence score.

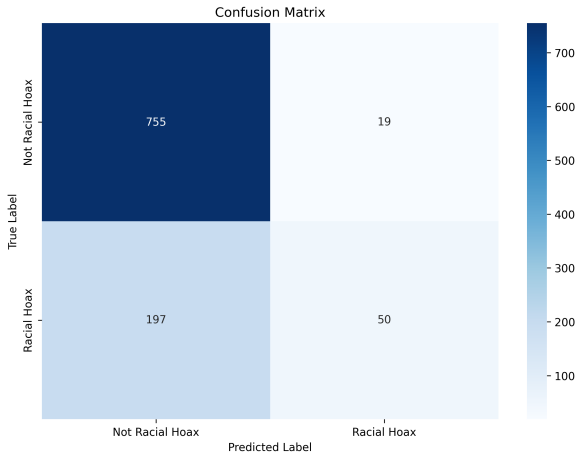

Figure 1: Confusion Matrix

4 Results

We compared several models for detecting hoax speech in code-mixed content. As shown in Table 1, our BERT-base model demonstrated the best overall performance on the development set, achieving an accuracy of 0.8000 and a macro-averaged F1-score of **0.6700**, outperforming XLM-RoBERTa (F1-score 0.5584), CNN (F1-score 0.5071), and Transformer-FFNN (F1-score 0.0863). Although XLM-RoBERTa achieved a slightly higher AUC score of 0.7781, BERT provided a better balance of precision (0.7300) and recall (0.6500), as well as a superior F1-score. On the official test set, our BERT model obtained an F1-score of **0.63**. These results emphasize the strength of multilingual-BERT for processing noisy, code-mixed data and demonstrate its real usefulness for multilingual hoax-speech detection

5 Discussion

Due to the linguistic complexity and social sensitivity involved, detecting racial hoaxes in code-mixed Hindi-English social media posts remains a challenging task. Our results demonstrate that multilingual transformer models, particularly BERT, perform well in this context. BERT achieved an F1-score of 0.6700 and an accuracy of 0.8000, reflecting strong generalization capabilities and effective contextual understanding, even when handling informal and noisy data. XLM-RoBERTa achieved the highest AUC (0.7781), reflecting good class separation, but its lower F1-score shows an imbalance between precision and recall. CNN, though faster and more efficient, lacked the depth to capture the nuanced meaning in racial hoax texts.

Model	Accuracy	Precision	Recall	F1-Score	AUC
mBERT	0.8000	0.7300	0.6500	**0.6700**	0.7720
XLM-RoBERTa	0.7818	0.5465	0.5709	0.5584	0.7781
CNN	0.7281	0.4511	0.5789	0.5071	0.7512
Transformer-FFNN	0.7508	0.3871	0.0486	0.0863	0.5692

Table 1: Model Comparison on the Development Dataset

Similarly, the Transformer-FFNN model underperformed, suggesting that shallow architectures struggle with the ambiguity and language mixing common in such posts. These results highlight the importance of deep contextual modeling for identifying deceptive narratives in multilingual environments. While BERT demonstrated the best overall performance, all models were challenged by code-switching and subtle sarcasm, highlighting the need for more diverse and culturally annotated training data. Table 1 illustrates how model depth and multilingual architecture influence performance.

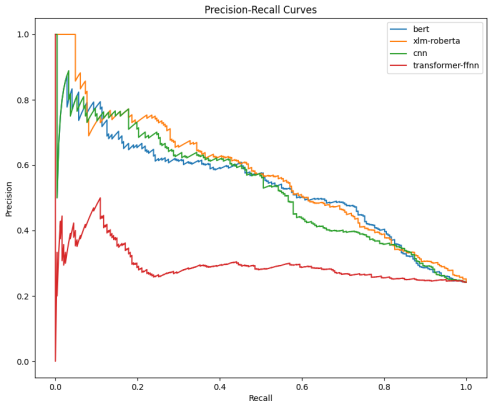

Figure 2: Precision and Recall plot on validation

6 Error Analysis

The confusion matrices in Figures 1 and 2 highlight a repeated pattern of misclassification, particularly for the minority class "Racial Hoax." Out of 247 actual "Racial Hoax" instances, 197 were misclassified as "Not Racial Hoax" indicating a strong bias toward the majority class. In contrast, the model performed well on the "Not Racial Hoax" class, correctly classifying 755 out of 774 cases. This imbalance indicates that the model has difficulty identifying the small linguistic or contextual signals that identify racial hoaxes. The results highlight the need to consider methods such as class balancing, deep semantic understanding, and advanced feature engineering. Increasing the sensitivity of the model about minority class characteristics would enhance the overall classification rate and reduce false negatives in critical categories such as racial hoaxes.

Conclusion

This work investigated the detection of racial hoaxes in Hindi-English code-mixed social media content using deep learning models. BERT outperformed other models in terms of F1-score, further demonstrating its ability to capture the contextual and linguistic nuances of bilingual, informal text. Despite its strong overall performance, the model struggled to correctly classify the minority class labeled as "Racial Hoax" showing a pronounced bias toward predicting the majority class. This indicates the persistent issue of class imbalance and the detection of delicate hints in minority classes. Remediation of this problem through class-aware methods and better representation of the features of the minority class will be the main way forward for future enhancements. Our work highlights the capabilities of multilingual transformers in code-mixed NLP, especially on socially sensitive tasks. Future studies should focus on tuning models with balanced datasets and incorporating richer semantic knowledge to improve the accurate identification of harmful or deceptive online content.

Limitations

While our work using BERT and other transformer-based models produced promising results in identifying racial hoaxes within code-mixed Hindi-English social media data, several limitations were observed. A major concern was the issue of class imbalance, which led the model to misclassify instances of the minority class and adversely affected its accuracy in detecting racial hoaxes. Besides, the data's mixed-code nature, usually involving informal language, transliteration, and non-uniform grammar, required more effort from the models, which were not adapted to such patterns. The ab-

sence of targeted pre-processing or code-mixed language modelling may have led to lower overall performance. The dataset used was relatively small and highly task-specific, limiting the generalizability of the results to broader, real-world scenarios. Furthermore, we have not yet used more sophisticated techniques like ensemble methods, data augmentation, or external knowledge integration, which could only increase the understanding of the model regarding complex, socially charged language.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Oluwatobi Joseph Abiola, Temitope Olasunkanmi Oladepo, Olumide Ebenezer Ojo, Grigori Sidorov, and Olga Kolesnikova. 2025a. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. pages 271–277.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025b. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. pages 262–270.

Tewodros Achamaleh, Tolulope Olalekan Abiola, Lemlem Eyob Kawo, Mikiyas Mebraihtu, and Grigori Sidorov. 2025a. Cic-nlp@ dravidianlangtech 2025: Detecting ai-generated product reviews in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 502–507.

Tewodros Achamaleh, Nida Hafeez, Mikiyas Mebraihtu, Fatima Uroosa, and Grigori Sidorov. 2025b. Cic-nlp@ dravidianlangtech 2025: Fake news detection in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 647–654.

Tewodros Achamaleh, Lemlem Kawo, Ildar Batyrshini, and Grigori Sidorov. 2024. Tewodros@ dravidianlangtech 2024: Hate speech recognition in telugu codemixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 96–100.

Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni. 2021. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Shankar Biradar, Kasu Sai Kartheek Reddy, Sunil Saumya, and Md Shad Akhtar. 2024. Proceedings of the 21st international conference on natural language processing (icon): Shared task on decoding fake narratives in spreading hateful stories (faux-hate). In *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*, pages 1–5.

Aashiq Bohra, Deepanway Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41.

Adriana C. Brisola and Ann Doyle. 2019. Critical information literacy as a path to resist "fake news": Understanding disinformation as the root problem. *Open Information Science*, 3(1):274–286.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Carlos Diaz Ruiz and Tomas Nilsson. 2023. Disinformation and echo chambers: how disinformation circulates on social media through identity-driven controversies. *Journal of public policy & marketing*, 42(1):18–35.

Lemlem Eyob, Tewodros Achamaleh, Muhammad Tayyab, Grigori Sidorov, and Ildar Batyrshin. 2024. Stress recognition in code-mixed social media texts using machine learning. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(1):32.

Don Fallis. 2015. What is disinformation? *Library trends*, 63(3):401–426.

Teri Finneman and Ryan J. Thomas. 2018. A family of falsehoods: Deception, media hoaxes and fake news. *Newspaper Research Journal*, 39(3):350–361.

Divina Frau-Meigs. 2019. Information disorders: Risks and opportunities for digital media and information literacy? *Medijske studije*, 10(19):10–28.

Muhammad Hatta. 2020. The spread of hoaxes and its legal consequences. *International Journal of Psychosocial Rehabilitation*, 24(03):1750–60.

Roland Imhoff and Pia Lamberty. 2020. A bioweapon or a hoax? the link between distinct conspiracy beliefs about the coronavirus disease (covid-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, 11(8):1110–1118.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Prashant Kapil and Asif Ekbal. 2024. A corpus of Hindi-English code-mixed posts to hate speech detection. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 79–85. NLP Association of India (NLPAI).

David M. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2017)*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Puneet Mathur, Ramit Sawhney, Maitreya Ayyar, and Rajiv Ratn Shah. 2018. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 11–17.

Sargun Nagpal, Sharad Dargan, Harsha Koneru, and Shikhar Rastogi. Innovations in code-mixed hate speech detection: The llm perspective.

Juan Manuel Pérez, Franco M Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and 1 others. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Rocco Scardigno, Adele Paparella, and Francesca D'Errico. 2023. Faking and conspiring about covid-19: A discursive approach. *The Qualitative Report*, 28:49–68.

Suman Si, Anupam Datta, Soumya Banerjee, and Sudip Kumar Naskar. 2019. Aggression detection on multilingual social media text. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Stephanie Jean Tsang. 2024. Misinformation, disinformation, and fake news? proposing a typology framework of false information. *Journalism*, page 14648849241304380.

Fida Ullah, Muhammad Zamir, Muhammad Arif, M. Ahmad, E. Felipe-Riveron, and Alexander Gelbukh. 2024. Fida@dravidianlangtech 2024: A novel approach to hate speech detection using distilbert-base-multilingual-cased. pages 85–90.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Dawei Yin, Zhiting Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the Web 2.0 (CAW2.0)*, pages 1–7.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

EM-26@LT-EDI 2025: Caste and Migration Hate Speech Detection in Tamil-English Code-Mixed Social Media Texts

Tewodros Achamaleh[1], Abiola T. O.[1], Mikiyas Mebiratu[2], Sara Getachew[2], Grigori Sidorov[1]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico
[2]Wolkite University, Department of Information Technology, Wolkite, Ethiopia
[2]Jimma University, Institute of Technology, Jimma, Ethiopia
Corr. email: sidorov@cic.ipn.mx

Abstract

In this paper, we describe the system developed by Team EM-26 for the Shared Task on Caste and Migration Hate Speech Detection at LT-EDI@LDK 2025. The task addresses the challenge of recognizing caste-based and migration-related hate speech in Tamil social media text, a language that is both nuanced and under-resourced for machine learning. To tackle this, we fine-tuned the multilingual transformer XLM-RoBERTa-Large on the provided training data, leveraging its cross-lingual strengths to detect both explicit and implicit hate speech. To improve performance, we applied social media–focused preprocessing techniques, including Tamil text normalization and noise removal. Our model achieved a macro F1-score of 0.6567 on the test set, highlighting the effectiveness of multilingual transformers for low-resource hate speech detection. Additionally, we discuss key challenges and errors in Tamil hate speech classification, which may guide future work toward building more ethical and inclusive AI systems.The source code is available on GitHub.[1]

1 Introduction

Increased hate speech on digital platforms constitutes a significant threat to the security of marginalised people. Harassment speeches are often targeted against people who associate themselves with some sensitive socio-cultural categories, like caste or migration status, in such highly culturally enriched states of India. The Tamil language, in its classical roots and its far-reaching use in South India and around the world, is widely used in social media, but is notably missing in many NLP datasets (Chakravarthi et al., 2020; Rajan et al., 2022). The urgent need to identify caste-based and migration-related hate speech in Tamil requires the development of culturally sensitive and technologically accurate automated tools (Mandl et al., 2019; Ranasinghe et al., 2023). To address this concern, LT-EDI@LDK 2025 launched the Shared Task on Caste and Migration Hate Speech Detection as its continuing effort to promote equality, diversity, and inclusion via language technologies (Rajiakodi et al., 2025). This work builds on past works on Tamil NLP, including shared tasks that have focused on speech recognition and accessibility for vulnerable people in Tamil-speaking communities (B. and others , et al.; Bharathi and others , et al.; Chakravarthi et al., 2021; Ramesh et al., 2021). Through this line of research, it becomes evident how necessary it is to develop language-specific approaches to ethical and inclusive AI (Blodgett et al., 2020; Joshi et al., 2020).

Therefore, EM-26 participated in creating a strong classification model that would be used to identify caste and migration-related hate speech in social media posts in the Tamil language. We chose the XLM-RoBERTa-Large multilingual transformer model because it promised to work on languages with little resources and morphology (Conneau et al., 2020a). Our strategy consisted of fine-tuning the XLM-RoBERTa-Large model with the organisers' labelled data to understand contextual subtleties in Tamil and hence identify the complex levels of hate speech existing. Our system illustrates the practical use of multilingual pre-trained models for socially essential tasks, even with limited resources, since it achieved a macro F1 Score of 0.6567 on the official test set. This paper details our approach, preprocessing steps, error analysis, and reflections on future improvements for caste and migration hate speech detection.

2 Related Work

The task of hate speech detection has picked up the pace in low-resource and culturally sensitive

[1]https://github.com/teddymas95/Caste-and-Migration-Hate-Speech.git

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 152–158

languages such as Tamil in recent years. As online hate attacks against caste and migration communities grow, making strong classification systems is both a technical and ethical necessity. First, rule-based approaches were not context-aware. Changing the workforce and long-term training are becoming major pains. This is what makes a shift to transformer-based models (e.g., XLM-RoBERTa (Conneau et al., 2020b), mBERT (Devlin et al., 2019)) so important. Ranasinghe and (Ranasinghe and Zampieri, 2021) showed that even for the case of zero-shot and few-shot scenarios, multilingual transfer learning was practical.

In the Indian setting, Chakravarthi and others (et al.) presented DravidianCodeMix for offensive and sentiment classification in Tamil-English and Malayalam-English. Subsequently, the Hope Speech Detection task (Chakravarthi and others , et al.) focused on socially inclusive content in Dravidian languages. These tasks demonstrated how performance is enhanced through fine-tuned transformers in the code-switched environment. The LT-EDI workshop series has been essential in developing underrepresented languages. (B. and others , et al.) and (Bharathi and others , et al.) works concerned inclusive technologies for vulnerable communities and HASOC sharing tasks (Mandl et al., 2021, 2023), they provided multilingual datasets for hate speech in the Indo-European and Dravidian languages. Specialist research has been conducted recently regarding hate speech among Tamils: (Senthilkumar et al., 2023) pointed out the esoteric nature of casteist language, and (Pandey et al., 2023) deep-published a fine-grained multilingual benchmark. Prompt-based learning (Roy et al., 2024) and contrastive learning (Velankar et al., 2023; Roy et al., 2024) and (Velankar et al., 2023) were promising.

(Achamaleh et al., 2024; Eyob et al., 2024) created a hate speech system for Telugu-English code-mixed text, extending the evidence for the effectiveness of transformer–based models in low-resource settings. The LT-EDI@LDK 2025 task extends the scope of this field by explicitly targeting caste and migration hate speech. Our method with XLM-RoBERTa-Large helps toward fairer and culture-attuned NLP solutions. Related efforts by the CIC-NLP team have demonstrated the effectiveness of multilingual transformer models in detecting AI-generated and deceptive content across diverse languages, including English and Dravidian code-mixed text (Abiola et al., 2025a,b; Achamaleh et al., 2025). These studies further support the applicability of transformer-based architectures such as XLM-RoBERTa in addressing socially sensitive and linguistically complex tasks like caste and migration hate speech detection.

3 Dataset Analysis

The dataset for our system includes annotated Tamil-English code-switched social media posts for caste and migration-related hate speech. The training set contains 5,512 samples, of which 3,415 are labeled non-hate speech (label 0) and 2,097 are labeled caste/migration-related hate speech (label 1). Such a distribution results in a modest class skew, with hate speech occurrences representing approximately 38% of the data. The development set has 787 samples, 485 of which are non-hate (label 0) and 302 of hate speech (label 1). While the dev set is somewhat balanced, the imbalance in training data necessitated additional strategies to ensure fair learning between classes. To compensate for this imbalance, and especially in the training, we exploited oversampling of the hate speech class and highly aggressive data augmentation. Through this, the present approach intends to expose the model to a broader set of linguistic variants related to hate speech while preserving diversification in the training examples. The use of label-aware augmentation and loss function finetuning assists the model to generalise better to minority-class cases that are crucial in real-world detection of hate speech.

4 System Overview

We use a binary classification approach in searching for caste and migration-related hate speech in Tamil-English CS posts on social media. We fine-tune the XLM-RoBERTa-Large transformer model, which is known for its powerful multilingual representation and is exceptionally efficient for low-resourced and code-mixed settings. The task entails labelling content as non-hate (0) or hate speech (1). To correct class imbalance and make better guesses on the instances of minority classes, we use focal loss instead of a regular cross-entropy with alteration of gamma and alpha parameters. On pre-processing, we clean missing values and normalize labels. To balance the dataset, we have oversampled for hate speech instances five times and augmented it with strong language-level augmentations, including synonym replacement, token dropout, and word scrambling, to diversify the use

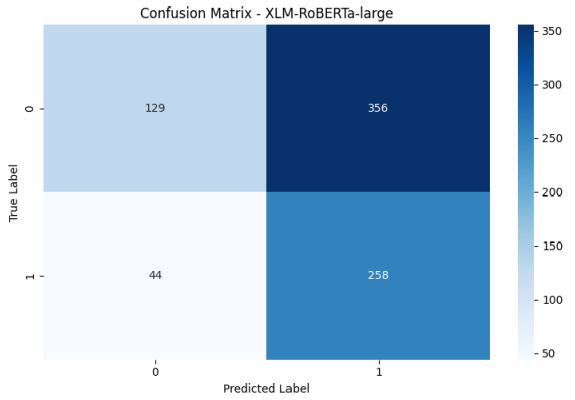

Figure 1: Confusion Matrix

of language. Text is to be tokenized by the multilingual tokenizer of the XLM-RoBERTa-Large, which would apply padding or truncation based on a max length of 128. The optimizer used to train the model is AdamW with gradient clipping and accumulation. We use a learning rate of 5e-6 with ReduceLROnPlateau and early stopping on validation F1 score. In order to enhance hate speech recall, we set the threshold for classification to 0.3 while inferring. Performance evaluation is done using precision, recall, weighted, and per-class F1 scores, and a confusion matrix. The best model checkpoint is auto-saved based on validation F1, also with the tokenizer and fine-tuned weights for easy deployment. Our system is designed for the LT-EDI@LDK 2025 task, addressing central issues of class imbalance, low-resource limitations, and code-switching, improving the detection of hate speech in challenging linguistic and social environments.

5 System Setup and Experiments

5.1 System Setup

Our system has been implemented in Python with PyTorch and Hugging Face Transformers. We fine-tune XLM-RoBERTa-Large for binary classification, benefiting from its power of multilingualism, particularly important for code-switched Tamil-English text. To counter the imbalance of labels, we apply Focal Loss ($\alpha = 0.9$, $\gamma = 3.0$) when we pay closer attention to the class of hate speech (label 1). The training and development data are brought to a status for training and development from CSV files with the help of pandas. In order to balance the dataset, we oversample samples of hate speech with a ratio of 5:1. For improvement, we use synonym substitution, keyword swapping, and token dropout with the help of nlpaug, predominantly aiming at hate speech examples. Tokenization is carried out by XLM-RobertaTokenizer, with a maximum length of 128 for the sequence with padding and truncation. We create a custom HateSpeechDataset, and we use PyTorch's DataLoader. Gradient accumulation (after every 4 steps) and gradient clipping are used for stability. Optimization is made with AdamW (learning rate 5e-6), and ReduceLROnPlateau carries out learning rate adjustment. In the training phase, training runs between 1 to 10 epochs, and a stop from the lack of progress of the validation F1 scores is triggered after 3 passes. In inference, hate speech's decision threshold is decreased to 0.3 to enhance recall. We evaluate using weighted and per-class F1 scores, a confusion matrix, and a classification report. The best-performing model and tokenizer are stored. GPU acceleration is utilized when available, as are logging and NLTK resources, for tracking and enhancement purposes.

5.2 Experiments

We evaluated the XLM-RoBERTa-Large model in the Tamil-English caste and migration hate speech dataset. The dataset was separated into a training set and a development set, and the hate speech class was very underrepresented. To address the class imbalance issues, we randomized and used the oversampling method to over-sample 5 times the minority class within the training set. Then we applied extensive textual augmentation methods, such as synonym insertions, word swaps, and random word deletion. In order to modify the model for the binary classification, Focal Loss was used with the values $\alpha = 0.9$ and $\gamma = 3.0$, which gave priority to learning from difficult and minority class samples. Training occurred for 10 epochs, using a learning rate of 5e-6 and batch size 4 with 4 steps of gradient accumulation. During training, the learning rate was adjusted dynamically according to the ReduceLROnPlateau scheduler based on changes in performance on the validation set, and early stopping occurred when the macro F1 score did not increase during the last three evaluations. We measured the model's effectiveness using weighted F1 Scores, class-wise F1 metrics, accuracy, and confusion matrices. To focus on finding hate speech in the minority class, a custom prediction threshold of 0.3 was used. The highest-performing model was retained and used on the validation dataset.

Model	F1 Score	Recall	Validation Loss
mBERT	0.3153	0.4295	0.0815
XLM-RoBERTa-base	0.2128	0.3837	0.0865
XLM-RoBERTa-Large	**0.4578**	**0.4917**	0.1039
CNN	0.2226	0.3863	0.0879

Table 1: Model Comparison on Validation Data.

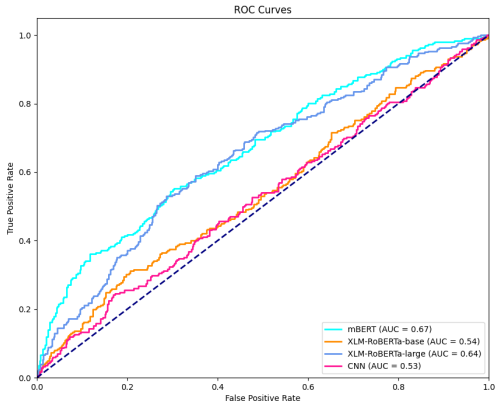

Figure 2: Roc Curves

6 Results

Our approach's score when used to evaluate the Tamil-English dataset used in detecting hate speech related to caste and migration was a weighted F1 Score of 0.6567. With XLM-RoBERTa-Large modeled with Focal Loss and aggressive augmentation, the model delivered impressive generalization in a scenario where resources are constrained and data is imbalanced. The class-wise model's results implied that it could reliably distinguish hate speech, and techniques such as oversampling, synonym-based augmentation, and threshold adjustment enhanced its performance. Although performance limitations, especially with respect to minority classes, are involved, the performance results provide grounds for advocating the use of class-focused learning and multilingual approaches for this task.

7 Discussion

Table 1 gives an overview of the F1 score of different models for the task of caste and migration-related hate speech detection. Of them, XLM-RoBERTa-Large had the highest performance on the validation set [F1 = 0.4578, recall = 0.4917] than smaller models such as mBERT, XLM-RoBERTa-base, as well as CNN. These results show that deeper transformer models that have strong language understanding in a variety of languages are better at catching hate speech in Tamil-English code-switched data. Although the large XLM-RoBERTa-Large model had a higher validation loss, it had better generalization when fed with Focal Loss and targeted oversampling. On the official test set, our best model managed a weighted 0.6567 F1 score, thus proving a favorable training strategy in a real-world, imbalanced setting. The increase in validation performance shows how augmentation, reduced thresholding, and class-aware loss led to increased recall of minority hate speech samples. As a whole, our approach demonstrates a key impact of prudent treatment of class imbalance, data augmentation, and model scaling on hate speech detection for under-resourced and socially sensitive settings such as Tamil-English. The performance gap between dev (F1 = 0.458) and test (F1 = 0.657) may raise concerns about data leakage. However, no overlapping samples or leakage patterns were found. The gap likely stems from class imbalance, oversampling, and threshold tuning that favored recall. We plan to use cross-validation and stricter data handling in future work.

7.1 Error Analysis

The confusion matrix shows that, when it comes to identifying 258 hate speech (true positives) and 129 true negatives of non-hate posts, the XLM-RoBERTa-Large model performed correctly. However, this model misclassified 356 instances of non-hate posts as hate speech (false positives). It is likely due to over-sampling, and the classification threshold is set to increase the recall at the expense of precision. The model only did not detected 44 instances of hate speech (false negatives), which demonstrates good recall performance. While this configuration guarantees most of the harmful content is tagged, this comes at the expense of speci-

ficity. The future improvements should address the reduction of false positives without affecting the recall, which is as high. This can be achieved by better threshold setting or more relevant data augmentation. Figures 1 and 2 show the confusion matrix.

Conclusion

Finally, our system is capable of solving the issue of caste and migration-related hate speech detection in Tamil-English code-switched social media content with the help of the XLM-RoBERTa-Large model. Using such advanced techniques as focal loss, aggressive data augmentation and threshold tuning, we accomplished a balanced performance, F1=$0.6567 on the test set. Despite strong recall by the model, particularly on the front of hate speech detection, the model records high false positives. This is a necessary tradeoff that is based on our priority to maximize harmful content coverage. The next step forward will consist of high-resolution classification thresholds, language-specific augmentation fine-tuning and integration of contextual clues to avoid misclassifications, with high recall quota. All-in-all, our approach advances the construction of responsible NLP systems in low-resource and culturally aware environments.

Limitations

In spite of the promising results achieved, our system has a number of limitations. First, the high false positives determined from the data's confusion matrix reveal that the classifier often misclassifies non-hate content as hate speech, and this the model may be getting influenced to identical linguistic patterns in code-switched Tamil-English data used to train the model. Second, while focal loss and oversampling ameliorated the problem of class imbalance, they probably made the model overfit for the minority class. Third, the use of synthetic data augmentation techniques (such as synonym replacement and token dropout) may lead to introducing noise or artificial variations that do not help generalize to the real-world inputs. In addition, the model can have difficulties in detecting the implicit or subtlest form of caste based hate, which depends on cultural and contextual understanding that goes beyond surface-level features. Finally, insufficiency in the availability of annotated data limits the model's capability to reflect the diversity of hate speech representations across the various dialects and sociolinguistic locations. Future studies should deal with these weaknesses in the form of larger datasets, context-aware modelling, and more intrinsic annotation guidelines. While we applied synonym replacement and token dropout to strengthen class 1 representation, we did not perform a formal ablation study to isolate their individual contributions. This remains a limitation, and future work will include controlled ablation to quantify the impact of each augmentation technique.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Oluwatobi Joseph Abiola, Temitope Olasunkanmi Oladepo, Olumide Ebenezer Ojo, Grigori Sidorov, and Olga Kolesnikova. 2025a. Cic-nlp at genai detection task 1: Leveraging distilbert for detecting machine-generated text in english. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 271–277.

Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide Ebenezer Ojo. 2025b. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 262–270.

Tewodros Achamaleh, Tolulope Olalekan Abiola, Lemlem Eyob Kawo, Mikiyas Mebraihtu, and Grigori Sidorov. 2025. Cic-nlp@ dravidianlangtech 2025: Detecting ai-generated product reviews in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 502–507.

Tewodros Achamaleh, Lemlem Kawo, Ildar Batyrshini, and Grigori Sidorov. 2024. Tewodros@ dravidianlangtech 2024: Hate speech recognition in telugu

codemixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 96–100.

Bharathi B. and others (et al.). 2022. Findings of the shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity and Inclusion*, LT-EDI 2022.

B. Bharathi and others (et al.). Overview of the fourth shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Joint SIGHUM and EACL 2025 Workshop on Language Technology for Equality, Diversity, and Inclusion*, LT-EDI 2025.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

B. R. Chakravarthi and others (et al.). 2021. DravidianCodeMix: Sentiment analysis and offensive language identification in code-mixed tamil-english. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 628–639.

B. R. Chakravarthi and others (et al.). 2022. Overview of the hope speech detection shared task for equality and inclusion. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity and Inclusion*, LT-EDI 2022, pages 139–148.

Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Ruba Priyadharshini, et al. 2021. Hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Sinnathamby Mahesan, and John P McCrae. 2020. A corpus for sentiment analysis of code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Armand Joulin, and Nicolas Usunier. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Lemlem Eyob, Tewodros Achamaleh, Muhammad Tayyab, Grigori Sidorov, and Ildar Batyrshin. 2024. Stress recognition in code-mixed social media texts using machine learning. *International Journal of Combinatorial Optimization Problems and Informatics*, 15(1):32.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Thomas Mandl, Pratik Modha, Pinkesh Badjatiya, and Aakash Bhatia. 2021. HASOC 2021: Hate speech and offensive content identification in multilingual contexts. In *Forum for Information Retrieval Evaluation*, FIRE 2021.

Thomas Mandl, Sanjay Modha, Punyajoy Majumder, Durgesh Patel, Mohana Dave, Chirag Mandlia, and Amit Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17.

Thomas Mandl, , Ekky P. Pamungkas, Ian Roberts, and . 2023. HASOC 2023: Overview of the hate speech detection subtask in indic languages. In *Forum for Information Retrieval Evaluation*, FIRE 2023.

Pushkar Pandey, Devansh Poonia, Abhinav Dwivedi, and Anupam Joshi. 2023. A multilingual benchmark dataset for hate speech detection in indian languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–125.

Vineeth Rajan, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, et al. 2022. Overview of the dravidianlangtech-2022 shared task on hope speech detection in dravidian languages. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI)*.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task on caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Gowtham Ramesh, Bhargav Murthy, Bharathi Raja Chakravarthi, et al. 2021. Classification of dravidian languages' offensive content using lstm and word2vec. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

Tharindu Ranasinghe, Constantin Orasan, and Marcos Zampieri. 2023. Multilingual hate speech and offensive language detection using cross-lingual language models. *Information Processing & Management*, 60(1):103207.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual hate speech detection with cross-lingual embeddings. In *Proceedings of the 13th International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1034–1041.

Aniruddha Roy, Arkadipta Bose, and Pinaki Bhattacharjee. 2024. Prompt-based multilingual hate speech detection in indic languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1300–1312.

Gokulakrishnan Senthilkumar, S. Stalin, and Natarajan Jegan. 2023. Annotated corpus for caste-based hate speech in tamil social media. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3783–3792.

Nikita Velankar, Apoorv Agarwal, Rajiv Ratn Sharma, and Pushpak Bhattacharyya. 2023. Contrastive learning for culturally aware hate speech detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11691–11705.

Hoax Terminators@LT-EDI 2025: CharBERT's dominance over LLM Models in the Detection of Racial Hoaxes in Code-Mixed Hindi-English Social Media Data

Abrar Hafiz Rabbani, Diganta Das Droba, Momtazul Arefin Labib
Samia Rahman, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology (CUET), Bangladesh
{u2004038, u2004064, u1904111}@student.cuet.ac.bd
u1904022@student.cuet.ac.bd , hasanmurad@cuet.ac.bd

Abstract

This paper presents our system for the LT-EDI 2025 Shared Task on Racial Hoax Detection, addressing the critical challenge of identifying racially charged misinformation in code-mixed Hindi-English (Hinglish) social media—a low-resource, linguistically complex domain with real-world impact. We adopt a two-pronged strategy, independently fine-tuning a transformer-based model and a large language model. CharBERT was optimized using Optuna, while XLM-RoBERTa and DistilBERT were fine-tuned for the classification task. FLAN-T5-base was fine-tuned with SMOTE-based oversampling, semantic-preserving back translation, and prompt engineering, whereas LLaMA was used solely for inference. Our preprocessing included Hinglish-specific normalization, noise reduction, sentiment-aware corrections and a custom weighted loss to emphasize the minority Hoax class. Despite using FLAN-T5-base due to resource limits, our models performed well. CharBERT achieved a macro F1 of 0.70 and FLAN-T5 followed at 0.69, both outperforming baselines like DistilBERT and LLaMA-3.2-1B. Our submission ranked 4th of 11 teams, underscoring the promise of our approach for scalable misinformation detection in code-switched contexts. Future work will explore larger LLMs, adversarial training and context-aware decoding.

1 Introduction

Racial hoaxes are harmful lies that falsely tie people or groups to crimes or events, often picking on specific ethnic or social communities to spark division or fear. The rise of hateful, racially charged speech—especially on platforms like Twitter and Facebook where users often blend languages like Hinglish (a mix of Hindi and English)—poses a serious challenge. The HoaxMixPlus dataset, consisting of 5,105 YouTube comments, serves as a key benchmark for detecting such harmful content.

Detecting racial hoaxes on social networks is challenging due to the difficulty in distinguishing truth from falsehood, the sheer volume of posts, and the intentional use of humor by users (Santoso et al., 2017). Traditional systems often misclassify content due to idioms, slang and subtle contextual cues. To address this, we leverage advanced models like the transformer-based CharBERT and LLM-based FLAN-T5[1], fine-tuned with task-specific instructions, rubrics, and prompt formulations. CharBERT's character-level embeddings and FLAN-T5's contextual understanding make them well-suited for interpreting nuanced, deceptive content.

This paper presents our submission to the Racial Hoax Detection Shared Task—a robust system leveraging transformers and LLMs, structured around three contributions:

- **Augmented Training for LLM and Transformer Models:** We trained the FLAN-T5 model on an augmented dataset generated using semantic-preserving back translation and SMOTE to mitigate class imbalance and enhance generalization in the low-resource setting. For the transformer-based CharBERT model, we applied an oversampling strategy to address class imbalance.
- **Class-sensitive training:** Introduced a weighted loss function in FLAN-T5 to increase sensitivity towards minority hoax instances and improve model fairness.
- **Transformer optimization:** To attain the best classification performance, the CharBERT model's hyperparameters were tuned using Optuna, an open-source framework for hyperparameter optimization.

Our approach demonstrates promising results in detecting racial hoaxes on code-switched social

[1] https://huggingface.co/google/flan-t5-base

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 159–170

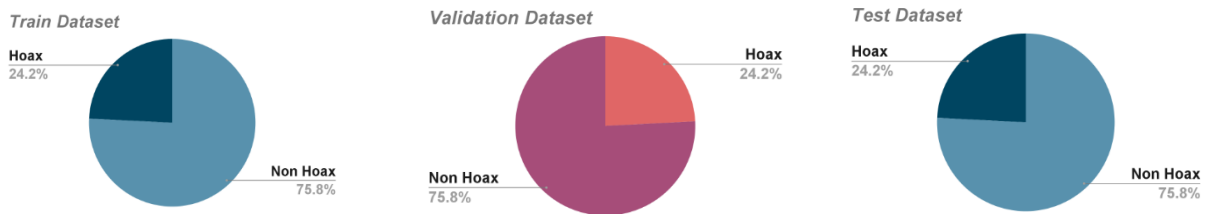

Figure 1: Dataset Distribution of Racial Hoax Dataset.

media platforms, highlighting the effectiveness of utilizing advanced transformer and LLM models alongside innovative data augmentation techniques like back translation, SMOTE and oversampling to tackle class imbalance and improve model generalization. For more details and to access the codebase, visit our project repository: GitHub Repository[2].

2 Related Work

In recent years, the detection of racial stereotypes and hoaxes in social media has become a critical focus of research. (Bosco et al., 2023) introduced a method for detecting racial stereotypes in Italian social media, focusing on the intersection of psychology and natural language processing (NLP). Building on these studies, (Schmeisser-Nieto et al., 2024) presented Stereohoax, a multilingual corpus annotated for racial hoaxes and stereotypes. Their work addresses a significant gap in understanding social media reactions to racial hoaxes and offers a valuable resource for future research. (Raza et al., 2024) explored the effectiveness of BERT-like models and large language models (LLMs) in the detection of fake news, focusing on the challenges posed by generative AI-annotated data. Their comparative evaluation provides a useful perspective on how different models perform in fake news and racial hoax detection tasks. (Banerjee et al., 2021) investigated transformer-based models for identifying hate speech and offensive content in both English and Indo-Aryan languages. Their work highlights the effectiveness of transformer models in multilingual environments, particularly in identifying harmful content in social media posts. (Guo et al., 2024) conducted a large-scale study on the use of LLMs for hate speech detection, focusing on the role of prompt engineering in improving the models' contextual understanding. Their findings suggest that LLMs can surpass traditional machine learning models in detecting hate speech when properly prompted. Recent work by (Carpenter et al., 2024) shows the effectiveness of fine-tuned FLAN-T5 models in educational settings, supporting our choice of FLAN-T5 for detecting racial hoaxes in code-mixed Hinglish. To address the issue of class imbalance, which is particularly critical in hoax detection where hoax instances are typically underrepresented, we draw inspiration from the dynamically weighted balanced (DWB) loss function proposed by (Fernando and Tsokos, 2021), which adaptively adjusts loss contributions based on class frequency and prediction confidence, enabling the model to focus on harder minority-class samples and improving generalization in imbalanced settings. Additionally, (Chakravarthi et al., 2025) presented an overview of the shared task on detecting racial hoaxes in code-mixed Hindi-English social media data, further advancing the understanding of racial hoaxes in multilingual contexts.

3 Dataset Description

The dataset (**?**) used in this shared task targets the challenge of racial hoax detection in Hinglish social media posts. It comprises real-world, user-generated content labeled with binary annotations: a label of 1 signifies the presence of a racial hoax, while 0 indicates a non-hoax instance. However, the dataset is notably imbalanced, with a significantly higher number of non-hoax examples. As illustrated in the pie charts of Figure 1, the training set contains 2,319 non-hoax cases versus only 741 hoax cases. The validation and test sets each contain 774 non-hoax and 247 hoax samples, creating an imbalance that can trip up standard classification models. These models often lean toward the majority class, making it tough to properly learn from the smaller hoax class. To tackle this, we applied methods like oversampling, SMOTE and loss function adjustments to better emphasize the minority class and enhance the model's effective-

[2] https://github.com/abrar-431/racial-hoax-detection-shared-task

ness.

4 System

In this section, we describe the methodologies employed for detecting racial hoaxes in social media content using two distinct approaches: Transformer Models and Large Language Models (LLMs).

4.1 Transformer Based Approach

Three transformer-based models were used in this study with the detailed illustrated in Figure 2 to detect racial hoaxes in Hinglish social media content.

CharBERT is used here which uses character-level embeddings to capture morphological subtleties and spelling variations in languages such as Hinglish which allow it to handle code-mixed, informal, and noisy text. This model works especially well for picking up on minute details in non-standard language usage, like that found in posts on social media [3].

DistilBERT is also used here which is a smaller and faster version of BERT. It was 60% faster and required fewer parameters while maintaining 97% of BERT's performance. This makes it perfect for real-time applications in large datasets.[4].

Finally, to address the multilingual nature of Hinglish, XLM-RoBERTa, a cross-lingual version of RoBERTa, was employed. It is proficient in understanding the contextual relationships between words in Hindi and English, having been trained on data from 100 languages. This makes it useful for cross-lingual tasks [5].

4.1.1 Data Preprocessing

The CharBERT tokenizer was used to perform tokenization because it is made to efficiently process the input data, . All characters were changed to lowercase and special characters, links, and unnecessary symbols were removed to normalize the text data. We also used oversampling techniques to address class imbalance and guarantee a balanced distribution of racial hoaxes and non-hoaxes in the training dataset. To ensure that both classes were equally represented, we resampled the dataset using RandomOverSampler [6] from the imbalanced-learn library.

[3] https://huggingface.co/imvladikon/charbert-bert-wiki
[4] https://huggingface.co/distilbert/distilbert-base-uncased
[5] https://huggingface.co/FacebookAI/xlm-roberta-base
[6] https://riverml.xyz/dev/api/imblearn/RandomOverSampler/

4.1.2 Model Architecture

The model used for this study is CharBERT, for which we have the best performance. It is a variant of the well-known BERT (Bidirectional Encoder Representations from Transformers) architecture.

- **Embedding Layer:** In CharBERT, word-level and character-level embeddings are included. The input text is first tokenized into subword units by the model. It then transforms each subword token into a dense representation that contains both the character-level embedding (which captures the morphology of words, including slang, informal language, and spelling variations) and the word-level embedding (from pre-trained word embeddings).

- **Character-Level Processing:** The emphasis on character-level information of CharBERT is the primary distinction between it and conventional BERT models. In order to identify subtle patterns in the text, such as misspellings, colloquial abbreviations, and new terms frequently found in code-mixed languages or social networks, CharBERT employs a convolutional layer.

- **Output Layer:** CharBERT model generates predictions for classification tasks by overlaying the transformer encoder with a dense output layer. Usually, a sigmoid activation function is used for binary classification tasks, or a softmax activation function for multi-class classification. To differentiate between racial hoaxes and non-hoaxes, CharBERT was optimized for binary classification in our situation.

- **Optimization:** The CharBERT model is adjusted using a cross-entropy loss function during training. To improve convergence, the Adam optimizer is used in conjunction with the learning rate scheduling to dynamically modify the learning rate.

4.1.3 Hyperparameter Optimization

To maximize the performance of transformer models, we employed an open-source hyperparameter optimization framework named Optuna [7]. A hyperparameter search space was established for learning rate, batch size, and the number of training

[7] https://optuna.org/

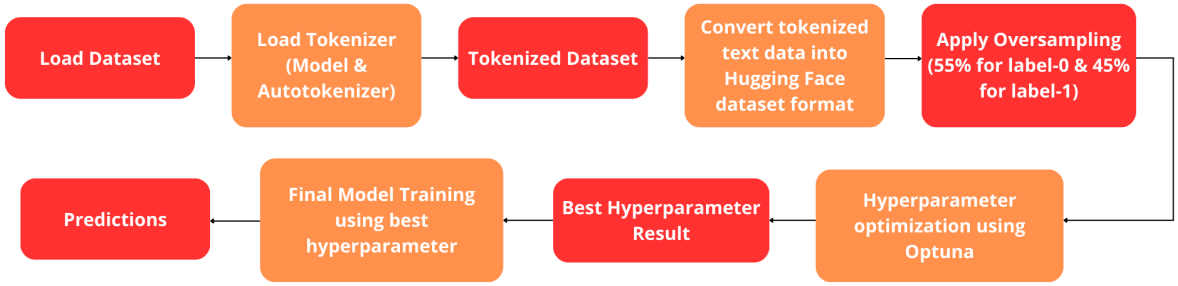

Figure 2: System flow for the Transformer-based approach.

epochs and weight decay. Maximizing the F1 score on the validation dataset served as the optimization's compass. The best model was chosen based on the evaluation outcomes of this optimization process.

4.1.4 Training

Trainer class from transformer's library was used for the training, and the recommended hyperparameters from Optuna were used. The F1 score, accuracy, precision, and recall metrics were used to monitor the model's performance, and early stopping callback was used to avoid overfitting.

4.1.5 Evaluation

The test dataset was used to assess the model following training. Performance metrics like accuracy, precision, recall, and F1 score were calculated by comparing the predictions with the true labels. The model's performance in both classes was thoroughly examined using the classification report.

4.2 Large Language Models (LLMs)

For the Large Language Models (LLMs), we employed the FLAN-T5-Base model as the primary model for racial hoax detection, with the detailed flow illustrated in Figure 3. Additionally, we utilized the Llama-3.2-1B[8] for inference to explore its capabilities in generating contextual responses. However, we primarily focused on FLAN-T5 due to its superior performance in handling code-switched text, better fine-tuning efficiency on our augmented dataset and robust generalization across diverse linguistic patterns, which were critical for detecting racial hoaxes effectively in our social media dataset.

[8] https://huggingface.co/meta-llama/Llama-3.2-1B

4.2.1 Data Augmentation

To enhance model generalization and address the significant class imbalance in the HoaxMixPlus dataset (75.8% Non-Hoax vs. 24.2% Hoax), we employed two complementary data augmentation techniques.

- **Back translation:** Back translation is a method used to generate new paraphrased samples by translating text to another language and then back to the original language. This process preserves the original meaning while altering the wording and structure.

  In our approach, we used pre-trained MarianMT models from Helsinki-NLP to translate sentences from English to Hindi[9] and then back from Hindi to English[10]. This augmentation was applied exclusively to samples labeled as Hoax (`label=1`). Using Hugging Face Transformers, we implemented a batched inference pipeline for efficient and consistent translation. The paraphrased sentences were then added to the dataset, effectively doubling the number of Hoax-labeled examples.

  This augmentation improves lexical and syntactic diversity, helping the model generalize better across different ways misinformation can be phrased. We monitored the process using the `tqdm`[11] progress bar and ensured reproducibility by shuffling the final dataset with a fixed random seed.

- **SMOTE:** While back translation introduced linguistic variety, it was not sufficient to fully address the class imbalance. Therefore, we

[9] https://huggingface.co/Helsinki-NLP/opus-mt-en-hi

[10] https://huggingface.co/Helsinki-NLP/opus-mt-hi-en

[11] https://tqdm.github.io/

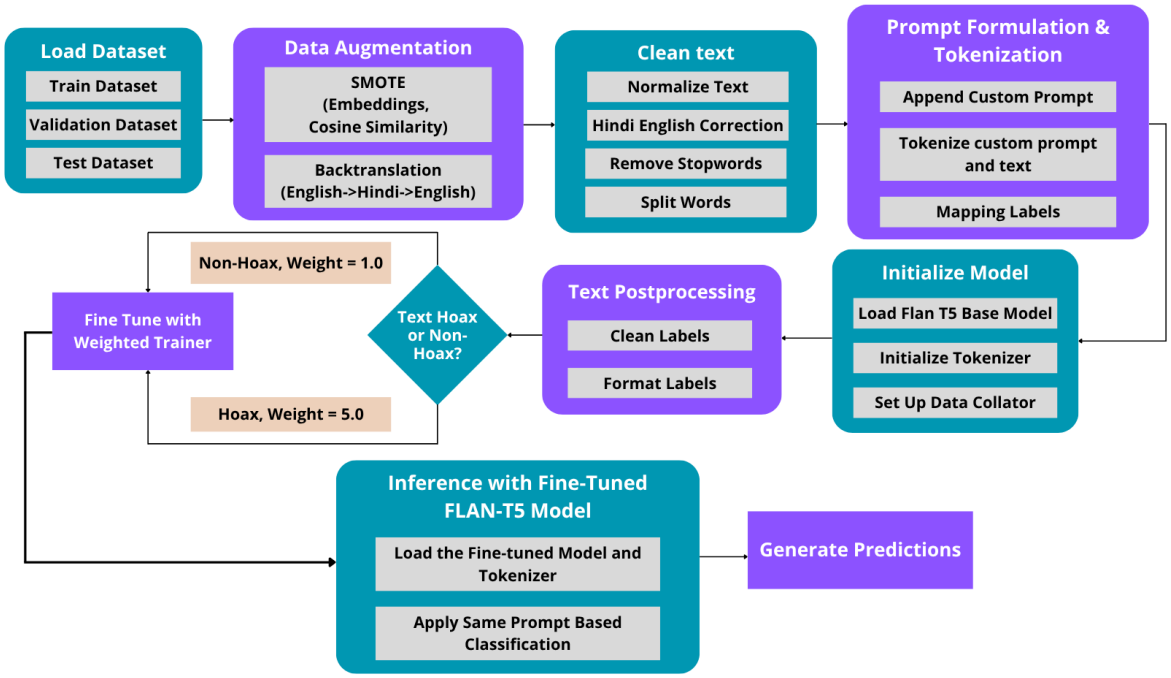

Figure 3: System flow for Fine-Tuning and Inference using the FLAN-T5 model.

also used SMOTE to synthetically generate new samples for the underrepresented Hoax class.

First, we encoded all text samples into vector representations (embeddings) using the SentenceTransformer model `all-MiniLM-L6-v2`[12]. SMOTE was then applied at the embedding level, generating new synthetic vectors by interpolating between existing Hoax samples. We set the oversampling ratio to 1.5 times the number of Non-Hoax samples to ensure balance.

To ensure that the generated vectors represented realistic content, we matched each synthetic vector to its closest original sentence using cosine similarity. This step helped maintain textual coherence in the generated samples.

In our LLM-based approach for detecting racial hoaxes on code-switched social media, we employed both back-translation and SMOTE to address the challenges of limited and imbalanced datasets. Back-translation enriched the dataset by generating diverse, semantically consistent variations of existing samples, preserving linguistic nuances critical for code-switched content. However, it alone was insufficient to handle severe class imbalances, as it primarily enhances sample diversity rather than balancing class distributions. SMOTE complemented this by synthetically generating samples for underrepresented classes, ensuring better representation of minority hoax categories. Using only one technique would have either limited diversity (with SMOTE alone) or failed to address class imbalance (with back-translation alone), compromising model performance. These augmented samples were consolidated into a Hugging Face Dataset, significantly improving class distribution and model robustness, enabling the LLM to generalize effectively across varied and imbalanced real-world scenarios.

4.2.2 Data Preprocessing

To handle the noisy, code-mixed nature of Hinglish social media text, we developed a custom preprocessing pipeline focused on normalization and token quality. We utilized textblob[13] for correcting English word fragments and estimating sentiment polarity where applicable. The pipeline involved lowercasing (while preserving sentiment-relevant punctuation), removal of numbers, URLs, emojis, and special characters. Hinglish-specific corrections were applied using a custom dictionary (e.g., "nhi" to "nahi", "pori" to "puri"), hybrid stopwords

[12] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[13] https://textblob.readthedocs.io/en/dev/

like "bhai" and "yar" were removed, and token fusion was addressed using regular expressions.

4.2.3 Model Architecture and Fine Tuning

We fine-tuned the FLAN-T5-Base model, a text-to-text transformer pretrained for instruction-following tasks, to perform binary classification of Hinglish social media posts into “Hoax” and “Non-Hoax” categories. Additionally, we leveraged the LLaMA-3.2-1B model for inference to evaluate its ability to generate contextual predictions, enhancing our exploration of LLM performance on the same dataset.

4.2.4 Prompt Engineering:

Various prompting strategies were explored—zero-shot, rubric-based, and few-shot (see Appendix A). Inputs followed the format: “Classify: <text>”, and outputs were generated as “0” (Non-Hoax) or “1” (Hoax), aligning with FLAN-T5’s instruction-tuned capabilities. To address severe class imbalance, a custom WeightedTrainer was implemented with a 5:1 loss weighting favoring the Hoax class, improving the model’s sensitivity to minority class instances. For the LLaMA-3.2-1B model, few-shot prompting was employed and proved most effective, leveraging its ability to adapt to contextual examples for improved inference performance.

4.2.5 Custom Weighted Trainer:

To deal with the strong imbalance between Hoax and Non-Hoax examples in the HoaxMixPlus dataset, we created a customized training approach that teaches the model to pay more attention to Hoax cases, which are much fewer in number. In a normal training setup, the model may become biased toward predicting the more frequent class (Non-Hoax) and ignore the less common but more important Hoax instances. To solve this, we made sure that the model gives more importance to correctly identifying Hoax examples by assigning a higher penalty when it gets them wrong. In simple terms, we told the model that misclassifying a Hoax is five times worse than misclassifying a Non-Hoax. This helped the model focus better on the minority class and significantly improved its ability to recognize misinformation, especially in real-world scenarios where such misleading content may appear less frequently but is more critical to detect.

4.2.6 Inference and Output Generation

Finally, predictions were generated using the best-performing checkpoint of the fine-tuned FLAN-T5-Base model, selected based on validation set performance. The test inputs were passed through the model in batches to ensure computational efficiency. Each output sequence generated by the model was decoded using the tokenizer to extract the predicted class labels, constrained to valid outputs—“0” representing Non-Hoax and “1” representing Hoax—to maintain label consistency. Additionally, for the LLaMA-3.2-1B model, inference was conducted using few-shot prompting, leveraging a small set of contextual examples to enhance prediction accuracy on the same dataset. The final, cleaned set of predictions was then compiled and stored in a structured CSV file format, enabling easy access for downstream evaluation, error analysis and comparison with other models. This completed a streamlined and effective end-to-end pipeline, encompassing training, evaluation, and inference.

5 Experiments and Results

In this section, we are presenting the experiments and results of our approaches-Transformer-based and Large Language Models for racial hoax detection in Hinglish social media text.

5.1 Experimental Setup

For the CharBERT model, we evaluated the performance on the HoaxMixPlus dataset, consisting of 5,105 Hinglish YouTube comments. The CharBERT model was fine-tuned for 20 epochs with a learning rate of $2e^{-5}$, weight decay of 0.01, and a batch size of 16. We used the RandomOverSampler technique used for addressing class imbalance to resample the dataset. Gradient accumulation with 4 steps was used to simulate a larger batch size, considering the memory constraints. Early stopping callback with a patience of 5 epochs was applied to prevent overfitting. The performance was monitored by using macro F1-score.

For the LLM (Flan-T5-base) model, we used a Kaggle P100 GPU to fine-tune the model on the same HoaxMixPlus dataset. The Flan-T5-base model was trained with a weighted loss function (5x for the Hoax class) to address class imbalance. Data augmentation techniques included back translation using MarianMT models[14]) and SMOTE.

[14] MarianMT Documentation

The training setup involved a learning rate of $5e^{-5}$, a batch size of 8, and early stopping based on macro F1 to select the best model checkpoint. These settings ensured that both models effectively handled the class imbalance and noisy, code-mixed nature of Hinglish data.

5.2 Parameter Setting

For the Transformer-based models, we initially used the CharBERT model and trained it for 20 epochs with a learning rate of $2e^{-5}$ and a weight decay of 0.01. The optimal learning rate schedule, warm-up ratio, and dropout values were selected automatically using Optuna framework to ensure the best hyperparameter configuration. A batch size of 16 was used, with gradient accumulation steps of 4 to simulate larger effective batch sizes, considering the memory constraints. To avoid overfitting, we applied early stopping callback with a patience of 5 epochs, monitoring the validation performance.

For the Large Language Models (LLMs), the model was trained for up to 10 epochs using the Adam optimizer, with a weight decay of 0.01 and a batch size of 8. The input sequences were truncated to 512 tokens, and early stopping was applied after 3 epochs without improvement in F1 score, ensuring that the best model checkpoint was selected. These settings were tailored to handle the complexity of detecting racial hoaxes in Hinglish social media data.

5.3 Evaluation Metrics

The macro F1 score, which balances precision and recall across both classes and is especially appropriate for our imbalanced binary classification task, was used to evaluate the performance of the Transformer-based models as well as the Large Language Models (LLMs). We provide a comprehensive classification report that includes precision, recall, and class-specific F1 scores in addition to the overall macro F1. This enables us to assess the model's ability to differentiate between hoax and non-hoax instances and identifies any remaining class-level flaws that might compromise the model's generalizability.

By concentrating on the model's advantages and disadvantages in identifying racial hoaxes, this method also aids in identifying differences in class performance in the case of LLM models. By disclosing these metrics, we hope to document the model's

5.4 Comparative Analysis

The performance of various classifiers across different model types is shown in Table 1. The results of our experiments with various transformer-based models and large language models (LLMs) reveal insightful performance trends. Among the transformer-based models, CharBERT achieved the highest macro F1 score of 0.70, alongside the highest accuracy (0.79), demonstrating its effective fine-tuning with Optuna for hyperparameter optimization. The DistilBERT models, both fine-tuned with Optuna and instructions, showed comparable performance with a macro F1 score of 0.66 and a weighted F1 of 0.77, indicating their strong performance despite being smaller variants of BERT. XLM-RoBERTa, another transformer model fine-tuned with Optuna, performed similarly to DistilBERT, with a macro F1 of 0.69 and a weighted F1 of 0.78, emphasizing its robustness for multilingual tasks.

When analyzing the performance of the FLAN-T5-Base and Llama-3.2-1B LLMs, the impact of different prompt variations becomes evident. FLAN-T5-Base, when fine-tuned with zero-shot prompting, achieved a macro F1 of 0.68. However, when fine-tuned with instructions, it showed an improved macro F1 of 0.67. The highest macro F1 score for FLAN-T5-Base was obtained when fine-tuned with Rubric, reaching 0.69. This highlights that different prompt strategies, including Rubric and few-shot prompting, can lead to varying results, with Rubric yielding the most optimal performance. On the other hand, Llama-3.2-1B, when used straight out of the box without any fine-tuning and just run in inference mode, struggled the most, hitting a low macro F1 of only 0.55. This really highlights why fine-tuning matters so much for large language models—tailoring them to specific tasks and carefully crafting prompts can make a huge difference in their performance.

To wrap it up, CharBERT and FLAN-T5-Base were the stars of the show. CharBERT delivered top-notch results with the highest macro F1 score when fine-tuned with Optuna, while FLAN-T5-Base, after being fine-tuned with the Rubric approach, achieved the best performance among the LLMs. This tells us that transformer-based models like CharBERT are strong contenders for this task, but models like FLAN-T5-Base can also excel with the right prompt tuning, especially when guided by strategies like Rubric.

Model Type	Model	Variation	F1 (Macro)	F1 (Weighted)
Transformer	**CharBERT**	**Fine-tuned with Optuna**	**0.70**	**0.78**
	DistilBERT Base	Fine-tuned with instructions	0.66	0.76
	DistilBERT	Fine-tuned with Optuna	0.66	0.77
	XLM-RoBERTa	Fine-tuned with Optuna	0.69	0.78
LLM	FLAN-T5-Base(248M)	Fine tuned with zero shot prompting	0.68	0.76
		Fine tuned with instruction	0.67	0.76
		Fine tuned with Rubric	0.69	0.79
		Fine tuned with Rubric & prompting	0.69	0.77
	Llama-3.2-1B	Inference	0.55	0.67

Table 1: Performance Evaluation of Different Models

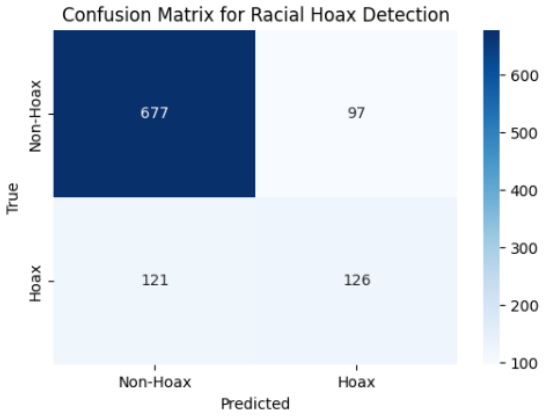

Figure 4: Confusion Matrix for CharBERT Model.

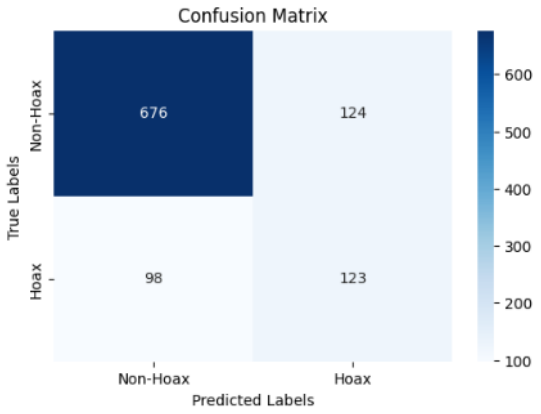

Figure 5: Confusion Matrix for LLM Model.

6 Error Analysis

In the error analysis, we evaluated the performance of the CharBERT model using the confusion matrix for racial hoax detection as shown in Figure 4. The CharBERT model correctly identified 677 non-hoax instances but misclassified 121 hoaxes as non-hoaxes and 126 non-hoaxes as hoaxes. These errors emphasize the need to reduce false negatives for better racial hoax detection. The confusion matrix in Figure 5 shows the confusion matrix with 676 true negatives, 124 false positives, 98 false negatives, and 123 true positives. These results highlight that the model performs well on Non-Hoax instances, with a solid precision of 87%, but struggles with Hoax classification, showing a lower precision of only 50%. This imbalance can be attributed to the class distribution, as Hoax examples are underrepresented in the dataset. The model tends to misclassify sarcastic posts as Non-Hoax (false positives) and hoaxes with neutral phrasing as Non-Hoax (false negatives). Augmentation strategies, such as back translation and SMOTE, helped mitigate some of these errors, but the challenge of distinguishing between subtle context variations remains. Additionally, short and noisy comments presented difficulties due to their inherent context ambiguity, further complicating accurate classification. In addition to the issues identified with sarcasm and neutral phrasing, the CharBERT model also faced difficulties when dealing with informal language and slang, common in Hinglish social media posts. This led to occasional misclassifications, especially when the context was subtle or ambiguous. Despite these challenges, the use of data augmentation techniques, like back translation and SMOTE, improved the model's performance by generating more diverse training examples. However, further improvements in handling noisy and short-text comments, along with enhancing the model's ability to detect nuanced hoaxes, will be necessary to address these shortcomings.

7 Conclusion

We introduced an innovative fine-tuned system for racial hoax detection in code-mixed Hindi-English YouTube comments. Specifically, we fine-tuned the CharBERT and Flan-T5 models on the HoaxMixPlus dataset, leveraging data augmentation techniques (including back translation and SMOTE), weighted loss optimization, and Hinglish-specific preprocessing to address challenges such as class imbalance, linguistic diversity, and contextual ambiguity. Our best-performing model, based on Flan-T5, achieved a macro F1

score of 0.69 on the test set, while the CharBERT-based model reached 0.70. These results underscore the effectiveness of integrating augmentation techniques with transformer-based architectures in low-resource, code-mixed settings. Future work will explore larger language models (LLMs), context-aware decoding tailored to Hinglish nuances, and constraint-based structured generation to further improve hoax specificity. Additionally, we plan to extend this research to other multilingual social media platforms and explore real-time detection mechanisms for dynamic hoax identification. Another promising direction involves fine-tuning models on even more diverse datasets to improve their generalization and robustness. Overall, this work demonstrates the potential of advanced NLP models in combating harmful misinformation in underrepresented languages and settings. Moreover, the findings highlight the importance of continued innovation in model architecture and training techniques to address the evolving nature of misinformation across diverse linguistic landscapes.

8 Limitations

Despite the promising results, several limitations emerged during our experiments. Data imbalance significantly impacted model performance, especially for the Hoax class. Although we employed a custom weighted loss function (with a 5:1 ratio) to prioritize hoax detection, both CharBERT and Flan-T5-base exhibited higher false negatives, indicating persistent challenges in recognizing hoax instances. CharBERT, while effective for non-hoax classification, struggled to generalize under imbalanced conditions.

Furthermore, the linguistic intricacies of code-mixed Hinglish presented challenges across both models. The Flan-T5-base model, in particular, showed sensitivity to subtle contextual shifts and the informal, slang-heavy nature of Hinglish. Our use of back translation, although beneficial for data augmentation, occasionally led to semantic drift, where paraphrased texts diverged slightly from their original meanings. Similarly, SMOTE generated synthetic samples that, due to their reliance on neighboring embeddings, often lacked linguistic diversity and richness, limiting their effectiveness in representing the true variability of hoax content.

Another key limitation stemmed from computational constraints due to restricted access to high-end GPU resources, we fine-tuned the Flan-T5-base variant rather than larger and more contextually expressive models like Flan-T5-large. This hardware limitation may have capped the model's capacity to capture deeper linguistic nuances and broader contextual signals.

Looking forward, future research should consider leveraging larger LLMs, integrating adversarial training to enhance model robustness against noisy and imbalanced data, and expanding Hinglish-specific lexicons to improve semantic understanding. Additionally, techniques like context-aware or constraint-based decoding could further enhance specificity in hoax detection by reducing ambiguity in model predictions.

References

Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. In *Proceedings of FIRE 2021: Forum for Information Retrieval Evaluation*. CEUR-WS.org. Presented at FIRE 2021, Virtual Event, 13th-17th December 2021.

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An italian social media corpus where psychology meets nlp. *Information Processing Management*, 60(1):103118.

Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, Mexico City, Mexico. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

K. Ruwani M. Fernando and Chris P. Tsokos. 2021. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951. Published in IEEE Transactions on Neural Networks and Learning Systems.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection. *arXiv preprint arXiv:2401.03346v1*. ArXiv:2401.03346v1 [cs.CY].

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2024. Fake news detection: Comparative evaluation of bert-like models and large language models with generative ai-annotated data. *arXiv*, 2412.14276v1. License: CC BY 4.0.

Irvan Santoso, Immanuel Yohansen, Nealson, Harco Leslie Hendric Spits Warnars, and Kiyota Hashimoto. 2017. Early investigation of proposed hoax detection for decreasing hoax in social media. In *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 123–128, Phuket, Thailand. IEEE.

Wolfgang S. Schmeisser-Nieto, Alessandra Teresa Cignarella, Tom Bourgeade, Simona Frenda, Alejandro Ariza-Casabona, Mario Laurent, Paolo Giovanni Cicirelli, Andrea Marra, Giuseppe Corbelli, Farah Benamara, Cristina Bosco, Véronique Moriceau, Marinella Paciello, Viviana Patti, Mariona Taulé, and Francesca D'Errico. 2024. Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, 2024:1–28.

Appendix A: Prompt Variations for Racial Hoax Detection

Prompt Type	Description	Prompt
Fine tuned with Zero Shot Prompting	A simple zero-shot prompt instructing the model to classify text as racial hoax (1) or not (0)	"Please check whether the text is racial hoax (1) or not (0):".
Fine tuned with Instructions	Instructs the model to classify Hinglish text as Hoax (1) or Non-Hoax (0) with a concise definition of both classes	"Classify the given Hinglish social media text as either 'Hoax' (1) or 'Non-Hoax' (0) for racial hoax detection. A Hoax (1) contains abusive, derogatory, or inflammatory language targeting a specific group (e.g., caste, religion, ethnicity), promotes hate or stereotypes, or includes threats or exaggerated claims to provoke fear or division. A Non-Hoax (0) is neutral, promotes unity, or discusses issues respectfully without targeting or dividing communities. Now classify this text:"
Fine tuned with Rubric	Provides a detailed rubric defining Hoax (1) and Non-Hoax (0) criteria, emphasizing strict classification for Hinglish text	"You are a binary text classifier. Classify the text strictly as Hoax (1) or Non-Hoax (0), prioritizing detection of Hoaxes. Follow this rubric tailored to Hinglish social media text: **Rubric:** - Hoax (1): Text is classified as Hoax if it: - Uses abusive, derogatory, or slang-heavy language (e.g., 'mule,' 'sale,' 'kute,' 'chamar') targeting a specific group (caste, religion, ethnicity, etc.). - Promotes hate, division, or stereotypes between communities (e.g., Hindu vs. Muslim, Dalit vs. Brahmin). - Contains threats (e.g., 'kat dalenge,' 'mita do') or exaggerated claims (e.g., conspiracies like 'gazwaehind') meant to provoke fear or anger. - Non-Hoax (0): Text is classified as Non-Hoax if it: - Encourages unity, respect, or neutral discussion across groups without hate. - Avoids abusive or inflammatory language, even if critical of issues (e.g., caste, reservation, politics). - Focuses on personal views, facts, or constructive critique without targeting or dividing communities. Now classify this text:"

Prompt Type	Description	Prompt
Fine tuned with Rubric + Few Shot	Extends the rubric-based prompt with few-shot examples from training data to guide classification of Hinglish text	"Classify the text strictly as Hoax (1) or Non-Hoax (0), prioritizing detection of Hoaxes. Follow this rubric tailored to Hinglish social media text: **Rubric:** - Hoax (1): Text is classified as Hoax if it: - Uses abusive, derogatory, or slang-heavy language (e.g., 'mule,' 'sale,' 'kute,' 'chamar') targeting a specific group (caste, religion, ethnicity, etc.). - Promotes hate, division, or stereotypes between communities (e.g., Hindu vs. Muslim, Dalit vs. Brahmin). - Contains threats (e.g., 'kat dalenge,' 'mita do') or exaggerated claims (e.g., conspiracies like 'gazwaehind') meant to provoke fear or anger. - Non-Hoax (0): Text is classified as Non-Hoax if it: - Encourages unity, respect, or neutral discussion across groups without hate. - Avoids abusive or inflammatory language, even if critical of issues (e.g., caste, reservation, politics). - Focuses on personal views, facts, or constructive critique without targeting or dividing communities. **Examples from Training Data:** 1. 'in suar ki aulad hone chali chali bache paida kar ke hindustan ka vada kar kar diya hai in a goli sa uda do' $\rightarrow$ 1 (Hoax: abusive ('suar ki aulad'), threatens violence, stereotypes population growth) 2. 'sikh hindu to ek hi hai hamara koi ghar aisa nahi hai yaha hindu sikh bacho ne shadi nahi ki ho' $\rightarrow$ 0 (Non-Hoax: promotes unity between Sikhs and Hindus, no hate) 3. 'quran or sariya kanon hi insaf kar sakta hai is duniya ka sabse mahan or prachin dharma sirf islam hai hidu to kide hai islam' $\rightarrow$ 1 (Hoax: derogatory ('kide'), promotes religious superiority and hate) 4. 'jab tak inter cast mariage ko society acept nahi karegi tab tak log kaise acept karenge ki reservation end ho' $\rightarrow$ 0 (Non-Hoax: neutral discussion on caste and reservation, no abuse or division) Now classify this text:"

CUET_Ignite@LT-EDI-2025: A Multimodal Transformer-Based Approach for Detecting Misogynistic Memes in Chinese Social Media

MD.Mahadi Rahman , Mohammad Minhaj Uddin , Mohammad Oman
and **Mohammad Shamsul Arefin**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904094, u1904118, u1904085}@student.cuet.ac.bd, sarefin@cuet.ac.bd

Abstract

Misogynistic content in memes on social media platforms poses a significant challenge for content moderation, particularly in languages like Chinese, where cultural nuances and multimodal elements complicate detection. Addressing this issue is critical for creating safer online environments, A shared task on multimodal misogyny identification in Chinese memes, organized by LT-EDI@LDK 2025, provided a curated dataset for this purpose. Since memes mix pictures and words, we used two smart tools: ResNet-50 to understand the images and Chinese RoBERTa to make sense of the text. The data set consisted of Chinese social media memes annotated with binary labels (Misogynistic and Non-Misogynistic), capturing explicit misogyny, implicit biases, and stereotypes. Our experiments demonstrated that ResNet-50 combined with Chinese RoBERTa achieved a macro F1 score of 0.91, placing second in the competition and underscoring its effectiveness in handling the complex interplay of text and visuals in Chinese memes. This research advances multimodal misogyny detection and contributes to natural language and vision processing for low-resource languages, particularly in combating gender-based abuse online.

1 Introduction

Misogynistic content fuels hostility and discrimination, particularly targeting women, and poses a significant barrier to fostering safe and inclusive online spaces. These memes flooding Chinese social media platforms like Weibo aren't just harmless jokes—they're digital barbs that mock women, blending snarky text with images to spread hostility (Kiela et al., 2020). Detecting misogyny in these multimodal formats is complex, as the intent hinges on the interplay between visual and textual elements (Chen and Pan, 2022). Subtle misogyny can dodge automated tools, or worse, those tools might flag innocent posts by mistake (Jindal et al., 2024). This is not just a technological problem, it is a social one, as these memes shape attitudes and amplify harm. The Misogynistic Meme Detection Shared Task at LT-EDI@2025 took aim at this, challenging teams to spot harmful memes in Chinese social media with precision. Our team, CUET_Ignite, participated in the LISN 2025 CodaLab competition to wrestle with these issues. We set out to build a system that could handle the tricky interplay of images and Chinese text. Our key contributions include the following:

- Used ResNet-50 to dig into images and Chinese RoBERTa to decode Chinese text, nailing the visual and linguistic cues of misogyny.
- Ran tests on image-only and text-only models to figure out which pulls more weight, landing an F1-score of 0.91 for solid accuracy and balance.

Inspired by (Rahman et al., 2025), which tackled abusive Tamil text with transformers, we pushed their ideas into the multimodal world of Chinese memes. This is our working way of making the Internet less toxic. For more details, our code is available at https://github.com/MHD094/Chinese-Misogyny-Meme-Detection.

2 Related Work

Social media is packed with harmful content such as misogyny and hate speech. In recent years, NLP researchers have been working on spotting trolled, hostility, and abusive content on social media. The early work was mostly about text alone (Anzovino et al., 2018). (Nozza et al., 2021) showed that hate speech tools struggle with different hate types, so misogyny needs its own focus.

Now, researchers are working on memes that mix text and images, making things trickier. Recent research has investigated multimodal approaches

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 171–176

to boost classification performance. For example, (H et al., 2024) developed a method to label Tamil and Malayalam memes as "Misogynistic" or "Non-Misogynistic" using Multinomial Naive Bayes, merging results with weighted probabilities. (Chen et al., 2024) used a CLIP model to see how text and pictures work together in misogynistic memes. (Mahesh et al., 2024) studied Tamil and Malayalam memes, pairing mBERT or MuRIL with ResNet-50, hitting F1-scores of 0.73 and 0.87. (Attanasio et al., 2022) built a Perceiver IO system, blending ViT for images and RoBERTa for text, which performed well at catching misogyny in memes. (Jha et al., 2024) launched MultiBullyEx, a dataset for cyberbullying memes in mixed languages. A Contrastive Language-Image Pretraining (CLIP) projection based multimodal shared-private multitask approach has been proposed there for visual and textual explanation of a meme. (Ahsan et al., 2024) shared MIMOSA, with 4,848 Bengali memes, using a fusion method to sort aggression. (Zhou et al., 2024) introduced Multi3Hate, a multilingual meme dataset capturing cultural variability in hate interpretation, and evaluated several vision-language models on this task. Similarly, (Lee et al., 2022) proposed Hate-CLIPper, which achieved state-of-the-art results by modeling cross-modal interactions between CLIP-encoded image and text features. These efforts show how hard it is to catch harmful memes in different languages and cultures, especially with subtle humor or jabs. Our work at CUET_Ignite@LT-EDI-2025 (Chakravarthi et al., 2025) stands out as we used ResNet-50 and Chinese RoBERTa to nab misogynistic Chinese memes, hitting an F1-score of 0.91. We addressed Chinese slang, idioms, and cultural vibes, making our model relevant for China's social media and helping to keep online spaces safer.

3 Task and Dataset Description

The pervasive spread of harmful content on social media, especially misogynistic material, has become increasingly common often hidden within memes that combine both text and images. These memes can reinforce negative stereotypes and promote gender-based hate speech. This work focuses on building automated systems that detect misogynistic memes by jointly analyzing visual and textual information, specifically in Chinese language memes. It is a multimodal classification task, where each meme must be categorized as either:

Misogynistic: Memes that contain content demeaning, targeting, or offending women.
Non-Misogynistic: Memes without harmful or offensive intent toward women.
The dataset requires analyzing both the image and the accompanying Chinese text, making the task challenging in the fields of Natural Language Processing and Computer Vision. It also contributes toward advancing multimodal and multilingual AI systems for hate speech detection.

This dataset builds upon the MDMD (Misogyny Detection Meme Dataset) originally introduced by (Ponnusamy et al., 2024), which focused on Tamil and Malayalam memes. The present dataset extends their methodology and annotation guidelines to Chinese social media content. A detailed overview of dataset design and objectives is also provided in (Chakravarthi et al., 2024). To ensure consistency with the original task, the same annotation schema was adopted and adapted for Chinese memes.

Classes	Train	Development	Test
Misogyny	349	47	104
Non-Misogyny	841	123	236
Total	1190	170	340

Table 1: Dataset distribution.

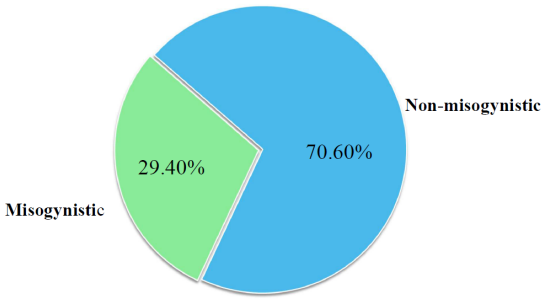

Figure 1: Percentage distribution of two different classes.

The dataset shows moderate imbalance, with 500 misogynistic and 1,200 non-misogynistic samples across all subsets. A total of 1200 memes are included: 1190 for training, 170 for development, and 340 for testing.

4 Methodology

The objective of this study is to detect misogynistic content in multimodal Chinese memes by integrating visual and textual features. Our approach begins with preprocessing the memes, followed

by feature extraction from both modalities. The features are then fused using an attention-based mechanism, and a classifier predicts whether the meme is misogynistic or non-misogynistic. Figure 2 provides a visualization of our methodology.

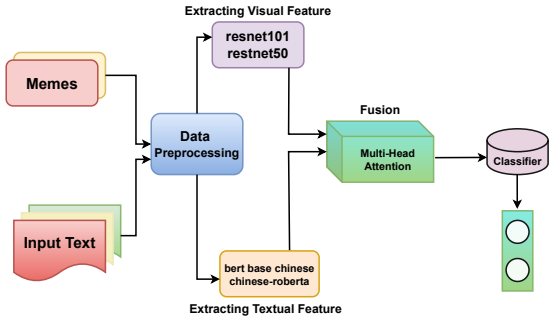

Figure 2: An abstract view of the proposed methodology

4.1 Data Preprocessing

In this step, we preprocess Chinese text and images for model compatibility. Text is tokenized using the hfl/chinese-roberta-wwm-ext tokenizer (Cui et al., 2020), transformed into 128-token numerical representations with [CLS] and [SEP] tokens, leveraging Chinese-RoBERTa's vocabulary for slang. Images are resized to 224×224 pixels, normalized with ImageNet statistics (Sayma et al., 2025), and converted to RGB.

4.2 Visual Approach

For visual feature extraction, we experiment with two pre-trained convolutional neural network (CNN) models: ResNet-50 and ResNet-101 (He et al., 2016), both pre-trained on ImageNet. The fully connected layer of each model is replaced with an identity layer to extract 2048-dimensional feature vectors. These models were chosen for their ability to capture complex visual patterns, such as culturally nuanced imagery or humorous elements in Chinese memes.

4.3 Textual Approach

The textual component of memes is processed using transformer-based models optimized for Chinese. We experiment with BERT-Base-Chinese and Chinese-RoBERTa-wwm-ext, both leveraging pre-trained weights. The [CLS] token's output from the last hidden state is extracted, generating 768-dimensional feature vectors. Chinese-RoBERTa-wwm-ext is prioritized for its whole-word masking strategy, which enhances its ability to capture contextual nuances in meme-specific language (Cui et al., 2021).

4.4 Multimodal Approach

Our multimodal approach combines the visual and textual features through a fusion strategy. The visual features from ResNet-50 or ResNet-101 (2048-dimensional) and textual features from Chinese-RoBERTa-wwm-ext or BERT-Base-Chinese (768-dimensional) are first projected to a common 512-dimensional space using linear layers, each followed by ReLU activation. These projected features are then fused using a multi-head attention mechanism (8 heads, embed dim=512) to capture cross-modal interactions, with the attention output averaged to produce a 512-dimensional representation (Wang et al., 2024). This combined representation is processed through a two-layer neural network classifier. The first layer reduces the dimensionality to 512, followed by ReLU activation and dropout (0.3) for regularization. The final layer produces binary classification outputs for misogyny detection. The training protocol uses Adam (learning rate: 1e-5, batch size: 16) for 10 epochs, with a weighted cross-entropy loss to address class imbalance, where class weights are computed as the inverse of class frequencies. This strategy aligns with recent approaches in multimodal harmful meme detection that emphasize cross-modal attention for enhanced feature fusion (Huang et al., 2024). Table 2 shows the list of tuned hyperparameters used in the experiment.

Hyperparameters	Value
Optimizer	Adam
Learning rate	1e-05
Epochs	10
Batch size	16
Dropout Rate	0.3

Table 2: Overview of optimized hyper-parameters.

5 Results & Discussion

This section presents a comparative performance analysis of various experimental approaches for classifying Chinese memes as misogynistic or non-misogynistic. The effectiveness is primarily assessed based on the weighted F1-score, while precision and recall are also considered in some cases. Table 3 presents a summary of the precision (P), recall (R), and F1 (F1) scores for each model on the test set. The results show that ResNet-50

and Chinese-RoBERTa-wwm-ext performed best among the visual and textual models, respectively, with an F1-score of 0.73 and 0.84. However, the top classification performance was observed in the multimodal models, where combining Chinese-RoBERTa-wwm-ext and ResNet-50 resulted in the highest F1-score of 0.91. These findings highlight the superiority of multimodal models in meme classification by effectively integrating text and visual features.

Approach	Classifier	P	R	F1
Visual	ResNet-101	0.70	0.72	0.71
	ResNet-50	0.73	0.74	0.73
Textual	Bert-Base-Chinese	0.76	0.77	0.77
	Chinese-Roberta	0.85	0.84	0.84
Multimodal	Bert-Base-Chinese + ResNet-101	0.83	0.86	0.84
	Chinese-Roberta + RestNet-50	**0.92**	**0.90**	**0.91**
	Bert-Base-Chinese + RestNet-50	0.88	0.83	0.85

Table 3: Evaluation of various models on the test set.

5.1 Quantitative Discussion

The results highlight the effectiveness of multimodal models in identifying misogynistic content in Chinese memes. The confusion matrix in Figure 3 shows that the multimodal model (Chinese-RoBERTa-wwm-ext + ResNet-50) outperforms unimodal approaches, correctly classifying 228 Not-Misogyny and 87 Misogyny instances, with fewer misclassifications (8 false positives and 17 false negatives). These findings affirm that leveraging both visual and textual features improves precision and recall in detecting misogynistic memes, particularly in reducing false positives.

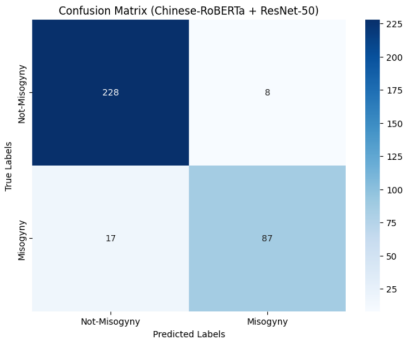

Figure 3: Confusion matrix of best performing approach.

5.2 Qualitative Discussion

Figure 4 presents sample predictions from our best-performing Chinese-RoBERTa-wwm-ext + ResNet-50 model. The first sample, incorrectly classified as non-misogynistic (label 0), contains

English Translation: cheap person

Actual: 1
Predicted: 0

English Translation: Baby, aren't you going to coax me to sleep?

Actual: 0
Predicted: 1

Figure 4: Examples of some misclassified samples from the top-performing model.

a derogatory term indicating misogyny (label 1), likely misclassified due to the model interpreting the cartoon character's mischievous expression as playful. Conversely, the second sample, genuinely non-misogynistic (label 0), was misclassified as misogynistic (label 1), possibly because the distressed cartoon cat's expression suggested conflict, despite the text's lighthearted tone. Cultural norms in Chinese internet memes, involving exaggerated expressions, may have influenced these errors. The multimodal model struggles with nuanced cases involving culturally specific language and ambiguous visuals, a challenge also seen in experiments with BERT-Base-Chinese and ResNet-101.

6 Conclusion

This work presented the details of the methods and performance analysis of the models for detecting misogynistic memes in Chinese, exploring visual, textual, and multimodal fusion techniques. The results revealed that the Chinese-RoBERTa-wwm-ext + ResNet-50 model achieved the highest F1-score of 0.91, demonstrating that multimodal fusion with attention mechanisms significantly enhances model performance. The attention-based fusion effectively captured cross-modal interactions, leading to improved precision and recall compared to unimodal approaches. In the future, we plan to explore advanced fusion strategies, such as cross-attention or graph-based methods, and extend the dataset to include more diverse meme content for better robustness, especially in handling Chinese internet slang and culturally specific references. Adding Chinese cultural knowledge and reducing model biases enhances adaptability, fairness, and generalization.

Limitations

A key limitation of this study stems from its dependence on pre-trained models for visual and textual feature extraction, which may not adequately address the intricacies of Chinese meme culture and context. Although the multimodal framework yields strong results, these models often lack the ability to generalize effectively to culturally specific or niche meme content. Moreover, the training dataset may not fully encompass the diverse range of Chinese memes, potentially undermining the model's robustness. The influence of cultural elements, such as humor, irony, and regional slang prevalent in Chinese online spaces, has also not been thoroughly examined. Misogynistic intent can often be conveyed indirectly through satire or cultural references, posing challenges for AI models in accurately discerning intent. Future efforts should focus on expanding the dataset, developing Chinese-specific models, and conducting in-depth analyses of humor and cultural influences to improve accuracy and adaptability. While the dataset was balanced and did not necessitate augmentation, applying data augmentation techniques in future work with larger, imbalanced datasets—through synthetic text or image transformations—could mitigate class imbalances and enhance generalization across diverse categories. This approach would lead to better performance in underrepresented scenarios, fostering a more resilient and effective model for practical deployment.

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, page 57–64, Berlin, Heidelberg. Springer-Verlag.

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022. MilaNLP at SemEval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 654–662, Seattle, United States. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025. An overview of the misogyny meme detection shared task for chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora Salim. 2024. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1864–1871, New York, NY, USA. Association for Computing Machinery.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17:e0274300.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Shaun H, Samyuktaa Sivakumar, Rohan R, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@LT-EDI 2024: A SVM-ResNet50 approach for multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226, St. Julian's, Malta. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 770–778, Las Vegas, NV, USA. IEEE.

Jianzhao Huang, Hongzhan Lin, Ziyan Liu, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with lmm agents. *arXiv preprint arXiv:2411.05383*.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jungbin Lee, Sungrae Jin, Donghyun Kim, Jihie Kim, and Jinwook Kim. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*.

Sidharth Mahesh, Sonith D, Gauthamraj Gauthamraj, Kavya G, Asha Hegde, and H Shashirekha. 2024. MUCS@LT-EDI-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287, St. Julian's, Malta. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

MD.Mahadi Rahman, Mohammad Minhaj Uddin, and Mohammad Shamsul Arefin. 2025. CUET_Ignite@DravidianLangTech 2025: Detection of abusive comments in Tamil text using transformer models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 392–397, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama, and Ashim Dey. 2025. CUET_Novice@DravidianLangTech 2025: A multimodal transformer-based approach for detecting misogynistic memes in Malayalam language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 472–477, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Xinyi Wang, Jun Liu, Min Zhang, and Wei Chen. 2024. Toxicn mm: A multimodal benchmark for chinese harmful meme detection. *arXiv preprint arXiv:2410.02378*.

Yujie Zhou, Yutai Ge, Qian Liu, Yue Zhang, and Paul Rottger. 2024. Multi3hate: Multimodal, multilingual, and multicultural hate speech detection with vision-language models. *arXiv preprint arXiv:2411.03888*.

girlsteam@LT-EDI-2025: Caste/Migration based hate speech Detection.

Towshin Hossain Tushi, Walisa Alam, Rehenuma Ilman, Samia Rahman
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u2004086, u2004015, u2004079, u1904022}@student.cuet.ac.bd

Abstract

The proliferation of caste- and migration-based hate speech on social media poses a significant challenge, particularly in low-resource languages like Tamil. This paper presents our approach to the LT-EDI@ACL 2025 shared task, addressing this issue through a hybrid transformer-based framework. We explore a range of Machine Learning (ML), Deep Learning (DL), and multilingual transformer models, culminating in a novel m-BERT+BiLSTM hybrid architecture. This model integrates contextual embeddings from m-BERT with lexical features from TF-IDF and FastText, feeding the enriched representations into a BiLSTM to capture bidirectional semantic dependencies. Empirical results demonstrate the superiority of this hybrid architecture, achieving a macro-F1 score of 0.76 on the test set and surpassing the performance of standalone models such as MuRIL and IndicBERT. These results affirm the effectiveness of hybrid multilingual models for hate speech detection in low-resource and culturally complex linguistic settings.

Keywords: Hate Speech Detection, Tamil, Code-Mixed Text, mBERT, BiLSTM, Caste, Tf-IDF.

1 Introduction

Hate speech encompasses expressions—verbal, written, or behavioral—that incite hostility or dehumanize individuals based on identity markers such as race, caste, gender, religion, or migration status. It reinforces prejudice and systemic discrimination, undermining individual dignity and social cohesion. Caste- and migration-based hate speech, in particular, reflects deep-rooted structural inequalities conveyed through derogatory narratives (Bhatt et al., 2022).

The rise of social media has amplified such harmful content, especially against marginalized communities in multilingual countries like India (Sharif et al., 2021). In this context, hate speech frequently appears in Tamil, English, and Tanglish—a code-mixed variant of Tamil in Latin script—posing unique challenges for automated detection. Tamil, a classical Dravidian language spoken by over 70 million people (Chakravarthi and Raja, 2020), presents significant challenges for Natural Language Processing (NLP) due to its complex morphology and limited annotated resources. Their workshop paper provided us an opportunity to engage with these challenges in processing mixed-up languages and to leverage our work (Rajiakodi et al., 2025). Recent advances in multilingual NLP have produced models tailored to Indian languages, such as IndicBERT, MuRIL, and mBERT (Khanuja et al., 2021). These transformer-based models, alongside machine learning (ML) and deep learning (DL) methods, form a robust foundation for tackling hate speech detection in such contexts. For the LT-EDI 2025 shared task, we propose a comprehensive system for identifying caste- and migration-related hate speech in Tamil social media. Our main objective was-

- To develop a robust multilingual system for detecting caste and migration related hate speech in Tamil social media by leveraging Tamil, English, and code-mixed (Tanglish) text.
- To propose a hybrid architecture that integrates transformer-based models (IndicBERT, mBERT, and MuRIL) with a BiLSTM network, combining contextual embeddings with sequential modeling to enhance classification performance in multilingual and code-mixed settings.

Our code, developed for this shared task can be accessed at [1]

[1] https://github.com/walisa810/Shared_Task_DravidianLangTech

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 177–182

2 Related work

Hate speech detection has gained growing attention, especially in multilingual and socio-culturally nuanced contexts. However, research focused on Tamil, particularly caste- and migration-related hate speech, remains sparse. Early work predominantly used traditional machine learning methods. For instance, Hossain et al. (2022) applied Logistic Regression to abusive Tamil text, and Bhimani et al. (2021) addressed caste and religion-based hate using similar approaches. In SemEval-2019 Task 5, Basile et al. (2019) and Almatarneh and Gamallo (2019) employed TF-IDF and lexicon-based features for multilingual hate speech detection. Sachdeva et al. (2021) used a Random Forest classifier for general hate speech classification. More recent studies have leveraged deep learning and contextual embeddings. Sharif and Hoque (2021) proposed an ensemble of CNN, BiLSTM, and GRU for Bengali hate speech, while Farooqi et al. (2021) combined Indic-BERT, XLM-R, and mBERT for code-mixed hate speech using conversational context. Romero-Vega et al. (2021) utilized SVM to detect xenophobic hate in Spanish tweets. Sajlan (2021) offered a qualitative analysis of caste-based hate speech, underscoring the need for robust computational approaches in this underrepresented area.

To address the complexities of caste and migration hate speech in Tamil, we propose a hybrid mBERT+BiLSTM model. This architecture combines the contextual understanding of mBERT with the sequential learning of BiLSTM to improve the detection accuracy in this underexplored domain.

3 Task and Dataset Description

This shared task focuses on **Caste and Migration Hate Speech Detection**. The objective of the task is to develop automatic classification models that can analyze social media texts, with a specific emphasis on content related to caste and migration. The task organizers provided a dataset containing social media posts in a mix of languages: The dataset comprises text in various language forms, including English (e.g., *"Mumbai Bangalore la 80 percentage outsiders"*), code-mixed English-Tamil, Tanglish (Tamil written in the Latin script), and pure Tamil.

Each entry in the dataset consists of the following:

- `id` – A unique identifier for the text
- `text` – The content of the text
- `label` – A binary class indicating presence of hate speech

The classification labels are defined as follows:

- **0** – Non Caste/Migration-related Hate Speech.
- **1** – Caste/Migration-related Hate Speech

The data set was segmented into training, development, and test data subsets to help with the thorough analysis and to facilitate model training. The class distribution is summarized in Table 1:

Table 1: Class distribution across datasets

Class	Train	Dev	Test
Non-Hate Speech (0)	3415	485	970
Hate Speech (1)	2097	302	606

4 Methodology

The growing prevalence of hate speech on social media has emerged as a critical issue, often targeting specific communities. In this section, we summarize the methods and strategies proposed to tackle the challenges highlighted earlier. Based on a thorough analysis, our research advocates for the adoption of a transformer-based model, using mBERT in combination with a BiLSTM architecture(Aodhora et al., 2025). Figure 2: presents a clear visualization of our methodology, illustrating the key steps in our approach.

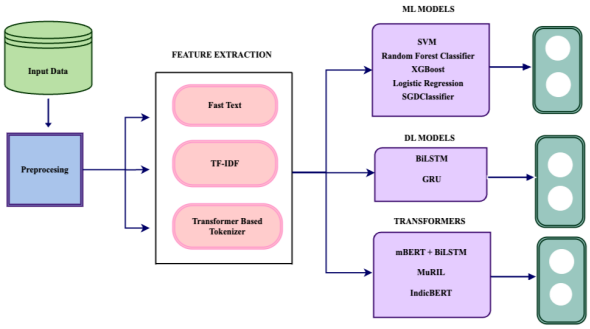

Figure 1: An abstract view of our methodology.

4.1 Data Preprocessing

The dataset provided by the problem organizing committee contains a significant amount of irrelevant and noisy content, including code-mixed

text (Ponnusamy et al., 2024). During preprocessing, we systematically cleaned the data by removing noise such as hyperlinks, emojis, punctuation, alphanumeric clutter, and special characters (e.g., slashes, brackets, and ampersands). In addition, all text was converted to lowercase to maintain consistency and improve data quality.

4.2 Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) was applied for feature extraction. In this approach, weights are assigned to words based on their frequency within a document and across the corpus, allowing the most informative terms to be identified. Furthermore, pre-trained Tamil FastText embeddings were utilized, which generate 300-dimensional word vectors by incorporating subword information through n-grams, thereby allowing the capture of semantic word relationships (Bojanowski et al., 2017).

For transformer-based models, we used model-specific tokenizers from the Hugging Face library to handle appropriate tokenization and padding, ensuring compatibility with each model's input requirements.

4.3 Model Building

In our research, we explored various machine learning (ML), deep learning (DL), and transformer-based models.

4.3.1 ML models

As a preliminary step, we evaluated the performance of several classical machine learning algorithms in our data set, including logistic regression, support vector machine (SVM), random forest, XGBoost, and stochastic gradient descent (SGD). These models served as baseline classifiers to assess the separability of the dataset and the inherent difficulty of the task. Overall, the models achieved moderate F1-scores ranging from 0.60 to 0.65, with SVM and XGBoost slightly outperforming others. These initial experiments provided a foundational benchmark for more advanced deep learning-based approaches.

4.3.2 DL models

Building on classical baselines, we developed an advanced deep learning pipeline that fused TF-IDF, FastText. These hybrid inputs were processed through sequential architectures, specifically Bidirectional LSTM and GRU networks, to capture long-range dependencies and contextual semantics in Tamil-English code-mixed hate speech. The use of BiLSTM was especially enabling it to comprehend contextual cues from both preceding and succeeding tokens—an essential capability for disambiguating morphologically rich code-mixed expressions. The BiLSTM model achieved a macro-averaged F1-score of 0.67, with the BiGRU yielding similar results, indicating consistent performance across architectures. These deep models outperformed classical approaches in recall and semantic awareness, establishing a strong benchmark for subsequent transformer-based exploration.

4.3.3 Transformer-based Models

The transformer-based models we applied include MuRIL, mBERT (Yu et al., 2024), and IndicBERT (Kakwani et al., 2020). These models were fine-tuned using their respective transformer-specific tokenizers to efficiently handle multilingual text. Table 2 presents the hyperparameters used across these models.

Among all, the mBERT + BiLSTM hybrid model demonstrated the best performance. In this architecture, we integrate the contextual strength of multilingual BERT (mBERT) with the sequential learning capability of a Bidirectional LSTM (BiLSTM). mBERT, pre-trained on over 100 languages, generates high-dimensional contextual embeddings (logits) from the input text, making it highly suitable for code-mixed and multilingual data such as Tamil-English (Tanglish) content.

These embeddings are concatenated with lexical features obtained from TF-IDF and FastText, forming a rich and diverse feature representation. This combined feature vector is reshaped and passed into a stacked BiLSTM network, which includes dropout and batch normalization layers for regularization and to prevent overfitting.We extracted logits as output features from all three models—m-BERT, MuRIL, and Indic-BERT. Token lengths of 128 and 512 were used with a batch size of 32 during feature extraction.

BiLSTM is particularly beneficial in this setup as it processes input sequences in both forward and backward directions, enabling it to capture long-range dependencies and contextual relationships that might be missed by unidirectional models. This is especially important for noisy, user-generated content, where the order and surrounding context of words can significantly impact meaning.

The fusion of mBERT's deep multilingual con-

textual embeddings with BiLSTM's sequential modeling allowed the architecture to effectively learn nuanced patterns in code-mixed text. This contributed to its superior performance, demonstrating its robustness and adaptability in multilingual hate speech detection tasks.

5 Result

In this section, we compare the performance of various machine learning (ML) and deep learning (DL) models. The effectiveness of each model is primarily evaluated using the macro F1-score. The hyperparameters for the DL models were manually fine-tuned based on their performance on the validation dataset. The sum of the precision (P), recall (R) and macro-F1 (MF1) scores for each model in the test set is presented in the Table 3

Table 2: The hyperparameters in BiLSTM model

Hyperparameters	Values
Input Shape	(1, 768)
Units	[256, 128, 64]
Dropout Rate	0.3
Optimizer	Adam
Loss Function	Binary Crossentropy
Batch Size	32
Epochs	100
Early Stopping	Patience = 5
Class Weights	Balanced (computed)

Table 3: Performance Comparison of All Classifiers

Classifier	P	R	MF1
SVM	0.65	0.65	0.65
RF	0.63	0.63	0.63
XGBoost	0.65	0.65	0.65
LR	0.64	0.64	0.64
SGD	0.64	0.64	0.64
BiLSTM	0.69	0.67	0.67
GRU + Attention	0.66	0.66	0.66
Muril	0.74	0.70	0.71
IndicBERT	0.65	0.65	0.64
mBERT + BiLSTM	**0.76**	**0.76**	**0.76**

We found that the m-BERT+BiLSTM model achieved the highest macro-F1 score of 0.76 on the test dataset using FastText embeddings, outperforming other machine learning and deep learning models. The combination of TF-IDF and FastText provided improved contextual understanding, while BiLSTM effectively captured sequential patterns in the data.

6 Error Analysis

6.1 Quantitative Discussion

To evaluate model performance, both quantitative and qualitative analyses were conducted. The confusion matrix Figure 2 showed 395 true negatives, 211 true positives, 90 false positives, and 91 false negatives, indicating a slight bias toward the negative class. Qualitatively, the mBERT+BiLSTM model effectively detected explicit caste- and migration-related hate speech in Tamil-English code-mixed text but struggled with satirical or metaphorical expressions. Errors were often due to class imbalance and subtle language use, highlighting the need for sociocultural and discourse-level understanding.

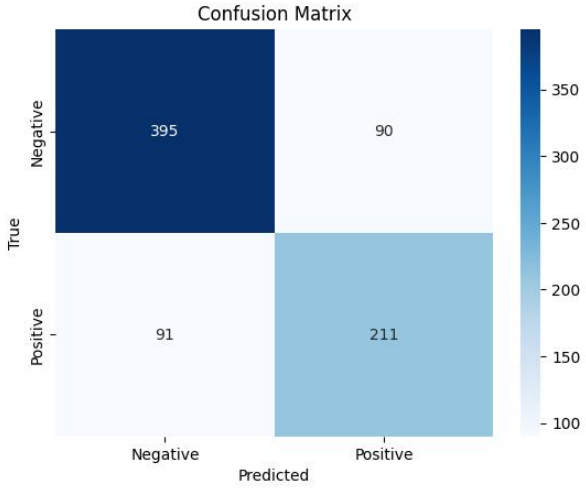

Figure 2: Confusion matrix of proposed model

7 Conclusion and Future Work

This work addressed the nuanced challenge of caste- and migration-based hate speech detection in Tamil by proposing a hybrid multilingual architecture. Integrating m-BERT's contextual embeddings with BiLSTM sequential modeling, and enriched by TF-IDF and FastText lexical features, our model achieved a macro-F1 score of 0.76, outperforming standalone baselines and affirming the utility of hybrid approaches in low-resource code-mixed settings. Looking ahead, future research can benefit from advanced ensemble techniques, such as weighted fusion of various transformer outputs or attention-based integration. Domain-adaptive pretraining on culturally grounded hate speech corpora

also holds promise for further boosting generalization. As the landscape of online discourse evolves, continuous innovation in modeling and data curation will be crucial in fostering safer, more inclusive digital environments.

Limitations

Despite the effectiveness of the mBERT+BiLSTM hybrid model in capturing contextual information, it struggled with nuanced and implicit hate speech, particularly in caste- and migration-related contexts. The training data set exhibited significant class imbalance, which led to biased learning and limited performance in minority classes. Furthermore, the multilingual and code-mixed nature of the data—especially Tanglish samples characterized by informal grammar, inconsistent spelling, and frequent code switching—posed considerable challenges. Data augmentation techniques such as synonym replacement and back translation were applied to address low-resource samples, but often introduced semantic noise due to transliteration inconsistencies, ultimately reducing the F1-score. The translation of Tamil text was also not feasible due to the lack of reliable parallel data. Furthermore, the model showed limitations in handling sarcasm, implicit hate, and vague contexts, which require deeper semantic and pragmatic understanding. Future work may explore transliteration-aware pretraining, more robust augmentation methods, and the use of pragmatic cues such as user metadata, comment threads, or conversational context.

References

Sattam Almatarneh and Pablo Gamallo. 2019. Citiuscole at semeval-2019 task 5: Combining linguistic features to identify hate speech against immigrants and women on multilingual tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 389–393, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sumaiya Rahman Aodhora, Shawly Ahsan, and Mohammed Moshiul Hoque. 2025. CUET_HateShield@NLU of Devanagari script languages 2025: Transformer-based hate speech detection in Devanagari script languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 260–266, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in nlp: The case of india. *arXiv preprint arXiv:2209.12226*.

Darsh Bhimani, Rutvi Bheda, Femin Dharamshi, Deepti Nikumbh, and Priyanka Abhyankar. 2021. Identification of hate speech using natural language processing and machine learning. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–4. IEEE.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Asoka Chakravarthi and Bharathi Raja. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Z. M. Farooqi, S. Ghosh, and R. R. Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

Alamgir Hossain, Mahathir Bishal, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. COMBATANT@TamilNLP-acl2022: Fine-grained categorization of abusive comments using logistic regression. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 221–228, Dublin, Ireland. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, and 1 others. 2021. Muril: Multilingual representations for indian languages. In *arXiv preprint arXiv:2103.10730*.

Rahul Ponnusamy, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sathiyaraj Thangasamy, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste/Immigration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Raúl R Romero-Vega, Oscar M Cumbicus-Pineda, Ruperto A López-Lapo, and Lisset A Neyra-Romero. 2021. Detecting xenophobic hate speech in spanish tweets against venezuelan immigrants in ecuador using natural language processing. In *Applied Technologies: Second International Conference, ICAT 2020, Quito, Ecuador, December 2–4, 2020, Proceedings 2*, pages 312–326. Springer.

Janak Sachdeva, Kushank Kumar Chaudhary, Harshit Madaan, and Priyanka Meel. 2021. Text based hate-speech analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 661–668.

Devanshu Sajlan. 2021. Hate speech against dalits on social media. *CASTE: A Global Journal on Social Exclusion*, 2(1):77–96.

Omar Sharif and Mohammed Moshiul Hoque. 2021. Align and conquer: An ensemble approach to classify aggressive texts from social media. In *2021 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pages 82–86.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Combating hostility: Covid-19 fake news and hostile post detection in social media. In *arXiv preprint arXiv:2101.03291*.

Boyang Yu, Fei Tang, Daji Ergu, Rui Zeng, Bo Ma, and Fangyao Liu. 2024. Efficient classification of malicious urls: M-bert—a modified bert variant for enhanced semantic understanding. *IEEE Access*, 12:13453–13468.

CUET_320@LT-EDI-2025: A Multimodal Approach for Misogyny Meme Detection in Chinese Social Media

Madiha Ahmed Chowdhury, Lamia Tasnim Khan, MD. SHAFIQUL HASAN, Ashim Dey
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004052, u2004048, u1904083}@student.cuet.ac.bd, ashim@cuet.ac.bd

Abstract

Detecting misogyny in memes is challenging due to their complex interplay of images and text that often disguise offensive content. Current AI models struggle with these cross-modal relationships and contain inherent biases. We tested multiple approaches for the Misogyny Meme Detection task at LT-EDI@LDK 2025: ChineseBERT, mBERT, and XLM-R for text; DenseNet, ResNet, and InceptionV3 for images. Our best-performing system fused fine-tuned ChineseBERT and DenseNet features, concatenating them before final classification through a fully connected network. This multimodal approach achieved a 0.93035 macro F1-score, winning **1st** place in the competition and demonstrating the effectiveness of our strategy for analyzing the subtle ways misogyny manifests in visual-textual content.

1 Introduction

The continuous rise of social media platforms has reshaped how people communicate and share content online. However, this digital transformation has also led to the proliferation of harmful content, including gender bias and sexism, often expressed through memes. Memes—though academically debated—have become deeply embedded in everyday digital interactions. Their combination of visual elements and textual content creates significant challenges for content moderation due to their multi-modal nature and the cultural context they carry (Chakravarthi et al., 2022a; Chakravarthi, 2022; Chakravarthi et al., 2022b). The *Shared Task on Misogyny Meme Detection in Chinese Social Media* at *DravidianLangTech@LDK 2025* aims to address this issue by identifying misogynistic content within Chinese memes. Chinese memes are unique in that they often include complex characters, cultural references, and idioms, necessitating specialized approaches for accurate detection. Given the limited research on multi-modal misogyny detection in Chinese social media, this task fills a significant gap in the field (Chakravarthi, 2020).

Our contribution to this shared task includes:

- Developing a novel multi-modal framework that effectively integrates visual and textual features from Chinese memes.
- Implementing specialized pre-processing techniques tailored for Chinese text and meme images to capture both linguistic and visual nuances.
- Demonstrating state-of-the-art performance using a fusion of `Chinese-BERT` and `Vision Transformer` models (Helboukkouri, 2020).
- Analyzing misogynistic patterns in Chinese memes to better understand the cultural nuances of online misogyny and how they manifest in this language.

Our approach achieved top performance in the task, advancing the field of multi-modal content moderation for Chinese social media and contributing to efforts aimed at creating safer online spaces. Our implementation details are available online[1].

2 Related Work

Detecting misogynistic content in memes presents unique challenges due to their multimodal nature. Research has approached this through text, images, or combined modalities.

Researchers in Pamungkas et al., 2020 have focused on text-based misogyny detection using traditional machine learning, primarily employing SVMs to identify hateful language. Other text-based approaches analyzed hate comments

[1] https://github.com/lamiatasnimkhan/CUET-320-Multimodal-Misogyny-Meme-Detection

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 183–188

to better understand linguistic patterns in misogynistic content (Tofa et al., 2025). However, text-only methods show limitations when targeting specific groups, highlighting the need for specialized misogyny detection datasets rather than generic hate speech systems. Similarly, vision-based research using CNNs and Transformers has demonstrated promise, but visual cues alone often prove insufficient. Integrating textual and visual features has yielded superior results. Systems combining Naive Bayes classifiers with visual processing have improved detection in low-resource languages, while advanced fusion techniques with Perceiver IO, RoBERTa, and Vision Transformers effectively address both binary and multi-label tasks (Pramanick et al., 2021). Other approaches use pre-trained CLIP models to bridge the semantic gap between modalities (Aho and Ullman, 1972). Recent work has extended multimodal misogyny detection to low-resource languages like Tamil and Tulu, emphasizing the need for culturally aware systems (Mallik et al., 2025).

3 Task and Dataset Description

We used the dataset from the Shared Task on Misogyny Meme Detection - LT-EDI@LDK 2025 (Ponnusamy et al., 2024; Chakravarthi et al., 2025), which includes 1,190 training, 170 development, and 340 test samples. The data comprises code-mixed Chinese-English memes, with a notable imbalance—fewer misogynistic samples than non-misogynistic ones (as shown in Table 1 and Figure 1). The dataset contains social media-style memes with real templates, combining sarcastic, code-mixed, and abusive captions with reaction images or symbolic visuals.

Set	Misogyny	Non-misogyny	Total
Train	349	841	1,190
Dev	47	123	170
Test	104	236	340

Table 1: Class distribution across dataset splits

Each transcription was annotated with a binary label:

- **Misogynistic:** Content that conveys hate, harassment, or derogatory views targeted at women.
- **Not Misogynistic:** Content that lacks misogynistic features or targets.

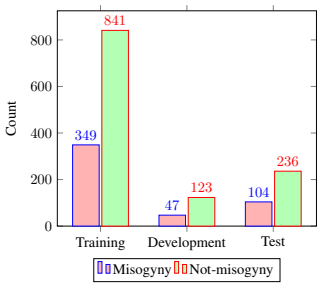

Figure 1: Dataset Distribution for Misogyny and Not-Misogyny

4 Methodology

The schematic representation of our approach is depicted in Figure 2.

4.1 Data Preprocessing

Meme samples underwent parallel preprocessing for both modalities. Text processing included regex-based URL and non-linguistic character removal, Jieba segmentation, filtering of 25 common Chinese stopwords, and truncation to 512 tokens for transformer compatibility. Images were validated, converted to RGB, resized to 224×224 pixels, and enhanced via histogram adjustment ($\alpha = 1.2$, $\beta = 20$). We implemented a dual loading strategy with OpenCV and PIL fallback, and normalized using ImageNet statistics.

4.2 Data Augmentation

Class-balanced augmentation is used exclusively on misogynistic samples (label=1) through three transformations: random brightness adjustment (factor=0.5), probabilistic grayscale conversion (p=0.2), and 4-bit color posterization. Each positive sample generated two augmented variants, effectively tripling the minority class representation. Text data remained unaugmented due to the semantic sensitivity of Chinese language transformations.

4.3 Feature Extraction

We employ separate pipelines for text and image feature extraction.

4.3.1 Text Modality

We employ three multilingual pretrained language models for text feature extraction. ChineseBERT (Cui et al., 2021) incorporates glyph and phonetic information for enhanced Chinese language processing. XLM-R (Conneau and Lample, 2019) utilizes the RoBERTa objective across 100 languages. The multilingual BERT baseline mBERT

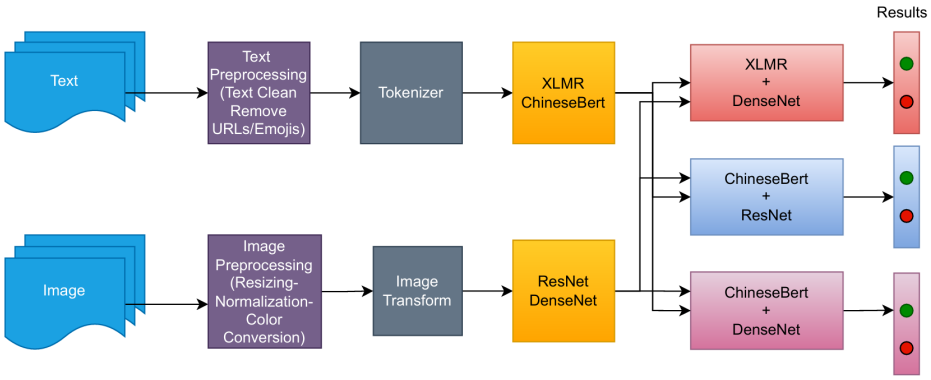

Figure 2: Abstract View of Methodology

(Devlin et al., 2019) covers 104 languages through masked language modeling. All models generate 768-dimensional embeddings via mean pooling of token outputs.

4.3.2 Image Modality

For visual feature extraction, we employ three CNN architectures: ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), and Inception-V3 (Szegedy et al., 2016). Each model's classification head is removed to extract 2048-dimensional features from ResNet-50 and Inception-V3, and 1024-dimensional features from DenseNet-121.

4.4 Multimodal Fusion

We combined our best unimodal models—ChineseBERT for text and DenseNet-121 for images—for multimodal fusion. Each modality was independently processed, with text encoded through mean-pooling of transformer token embeddings and visual data via flattened CNN output. These representations were then concatenated into a unified multimodal embedding. The fused embedding was processed through a fully connected layer with softmax activation. Rather than complex attention mechanisms, our simpler embedding concatenation approach reduced model complexity while maintaining strong performance, suggesting basic fusion suffices for this task. Table 2 illustrates the hyperparameters used for the models. All models used Adam optimizer.

5 Result and Analysis

Our experimental results demonstrate that multimodal fusion of ChineseBERT and DenseNet-121 achieves superior performance ($F1 = 0.93$) for misogyny meme detection, outperforming unimodal approaches where text-based ChineseBERT ($F1 = 0.91$) surpassed image-only models. The 2.2% improvement from multimodal integration confirms the complementary value of combining linguistic and visual features, with DenseNet-121 ($F1 = 0.82$) proving more effective than other CNN architectures for visual analysis, while maintaining balanced precision-recall ratios across all models. We evaluated the performance of our models using macro-averaged precision, recall and F1 score (Macro-F1). Among these, the Macro-F1 score serves as the primary metric for assessing the overall effectiveness of the systems.

Modality	LR	WD	BS	EP
Text	2×10^{-5}	0.01	16	10
Image	1×10^{-4}	0.00	16	20
Bimodal	1×10^{-4}	0.00	16	30

Table 2: Training hyperparameters for all models. LR: Learning Rate, WD: Weight Decay, BS: Batch Size, EP: Epochs.

5.1 Quantitative Analysis

We evaluated several models using macro F1-score (MF1) to identify the best approach for misogyny detection. Detailed result is presented in Table 3. Unimodal models served as baselines, while the multimodal DenseNet + Chinese BERT model achieved the highest MF1 of 0.93 by effectively leveraging both image and text features. As shown in Figure 3, this model correctly classified 227 non-misogynistic and 93 misogynistic samples, with 9 false positives and 11 false negatives. These errors likely stem from class imbalance and limited data diversity. Multimodal models are effective when misogyny arises from interactions between

text and images (e.g., sarcasm). Still, their performance may decline if one modality dominates or if fusion introduces noise. Image-only models may outperform due to noise in visual features, whereas text modalities often provide stronger discriminative signals; suboptimal fusion can degrade these signals. Following prior work, we trained unimodal models using joint labels. We note this may introduce label noise, as one modality may lack full context, and suggest future work explore modality-specific labels or weak supervision.

Type	Model	P	R	F1
Text	**ChineseBERT**	0.92	0.90	**0.91**
	XLM-R	0.91	0.88	0.89
	mBERT	0.89	0.87	0.88
Image	ResNet-50	0.79	0.75	0.76
	DenseNet-121	0.81	0.83	**0.82**
	InceptionV3	0.79	0.81	0.80
Multi-modal	ChineseBERT+ ResNet	0.84	0.78	0.80
	ChineseBERT+ DenseNet	0.94	0.92	**0.93**
	XLM-R+ DenseNet	0.85	0.77	0.79

Table 3: Macro-averaged classification scores: Precision (P), Recall (R), and F1 across model types.

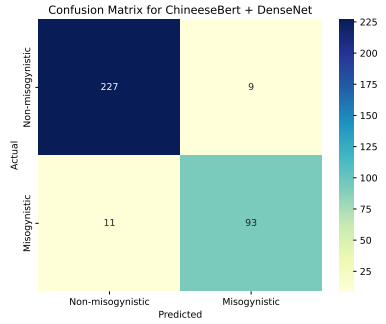

Figure 3: Confusion Matrix for Multimodal DenseNet + Chinese BERT Model.

5.2 Qualitative Analysis

Figure 4 highlights both correctly classified and misclassified cases. Among the misclassifications:

False Positives: Sample 882.jpg ("Friday Friday Reborn") was incorrectly flagged as misogynistic despite containing no gendered language, suggesting model overfitting on stylistic cues without context.

False Negatives: Sample 1562.jpg ("A handsome and smart baby boy invites you to take him home. Refuse to take him home.") went undetected despite containing objectifying language and gender stereotypes.

Image_ name	Result	Prediction	Transcriptions
1342.jpg	Misogyny	Misogyny	“你这么懒以后一定会被婆婆骂” “女生不用读太多书，反正最后都要嫁人的” “25岁还没对象我都替你丢人” “35岁之后就是嫁不出去的老姑娘了” ("You will definitely be scolded by your mother-in-law for being so lazy" "Girls don't need to study too much, they will get married in the end anyway" "I feel ashamed for you not having a partner at the age of 25" "After the age of 35, you will be an old maid who can't get married)
1582.jpg	Not-Misogyny	Not-Misogyny	光顾着上学 忘记上吊了(I was so busy with school that I forgot to hang myself.)
882.jpg	Not-Misogyny	Misogyny	周五周五 脱胎换骨 (Friday Friday Reborn)
1562.jpg	Misogyny	Not-Misogyny	帅气聪明的男宝宝 邀请你接他回家 拒绝 接他(A handsome and smart baby boy invites you to take him home. Refuse to take him home.)

Figure 4: Examples of the DenseNet + Chinese BERT model’s anticipated outputs with English translations.

These errors reveal the model’s difficulty with indirect misogyny, especially in sarcastic, metaphorical, or superficially neutral language. While it handles explicit discrimination well, detecting implicit bias and understanding cultural context requires further improvement.

6 Conclusion

In this study, we started by experimenting with a few unimodal models, focusing separately on text and image data. While these gave us a decent starting point, it was the multimodal models, which combine both visual and textual features, that really stood out. In particular, our DenseNet + Chinese BERT model achieved the best results, reaching an F1 score of 0.93. Despite working with a relatively limited dataset, these findings show that combining modalities is crucial for capturing the complex and often subtle nature of misogynistic content in memes. For future work, we plan to expand the dataset, explore better data augmentation, and fine-tune our multimodal fusion techniques to push performance even further.

Limitations

While our results are promising, model performance is limited by several constraints. The dataset size was relatively small, especially for detecting subtle or implicit misogyny. Despite data augmentation, exposure to diverse examples remains limited. Pretrained language models like Chinese-BERT and XLM-R may overlook nuances in slang, dialects, or sarcasm. Multimodal pairs—e.g., Chinese-BERT with ResNet or DenseNet—struggled with interpreting irony or misalignment between text and image, which is common in memes. Additionally, our current models do not support audio, leaving out video-based content for future work.

Ethics Statement

Our team conducted this research with a deep commitment to ethical standards. As researchers and internet users ourselves, we recognize the real harm that online misogyny causes to women and marginalized communities. By developing better methods to identify harmful content in Chinese memes, we hope our work contributes to creating digital spaces where everyone feels welcome and respected.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025. An overview of the misogyny meme detection shared task for chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

B.R. Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

B.R. Chakravarthi. 2022. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

B.R. Chakravarthi, R. Ponnusamy, and R. Priyadharshini. 2022a. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analysis*, 14(4):389–406.

B.R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, and 1 others. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity, and Inclusion*, pages 369–377.

Alexis Conneau and Guillaume Lample. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Zheng Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Houssam Helboukkouri. 2020. Characterbert: A pretrained language model using character-level inputs. `https://huggingface.co/helboukkouri/character-bert`. Accessed: 2025-05-13.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*, pages 4700–4708.

Arpita Mallik, Ratnajit Dhar, Udoy Das, Momtazul Arefin Labib, Samia Rahman, and Hasan Murad. 2025. CUET-823@DravidianLangTech 2025: Shared task on multimodal misogyny meme detection in Tamil language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 325–329, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Endang Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing Management*, 57:102360.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Soumick Pramanick, Dimitar Sharma, Dimitar Dimitrov, Prasenjit Mukherjee, Marcos Minovski, Marta Villegas Enrich, Miguel Angel García Carmona, Manuel Núñez-García, Javier Casas, Antti Ilari Vatanen, and Preslav Nakov. 2021. Detecting misogyny and xenophobia in Spanish tweets using multilingual contextual embeddings and multimodal information. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, pages 278–289. CEUR Workshop Proceedings.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826.

Farjana Alam Tofa, Khadiza Sultana Sayma, Md Osama, and Ashim Dey. 2025. CUET_Novice@DravidianLangTech 2025: Abusive comment detection in Malayalam text targeting women on social media using transformer-based models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 483–488, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Speech Personalization using Parameter Efficient Fine-Tuning for Nepali Speakers

Kiran Pantha[†], **Rupak Raj Ghimire**[†], and **Bal Krishna Bal**

Information and Language Processing Research Lab
Kathmandu University, Dhulikhel, Nepal
info@kiranpantha.com.np, rughimire@gmail.com, bal@ku.edu.np
[†]Equal contribution

Abstract

The performance of Automatic Speech Recognition (ASR) systems has improved significantly, driven by advancements in large-scale pre-trained models. However, adapting such models to low-resource languages such as Nepali is challenging due to the lack of labeled data and computational resources. Additionally, adapting the unique speech parameters of the speaker to a model is also a challenging task. Personalization helps to target the model to fit the particular speaker. This work investigates parameter-efficient fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA) and Decomposed Weight Low-Rank Adaptation (DoRA) to improve the performance of fine-tuned Whisper ASR models for Nepali ASR tasks by Personalization. These experiments demonstrate that the PEFT methods obtain competitive results while significantly reducing the number of trainable parameters compared to full fine-tuning. LoRA and DoRA show a relative WER to FT_{Base} increment of 34.93% and 36.79%, respectively, and a relative CER to FT_{Base} increment of 49.50% and 50.03%, respectively. Furthermore, the results highlight a 99.74% reduction in total training parameters.

1 Introduction

Automatic Speech Recognition (ASR) systems like voice assistants are widely used with the rapid development of deep learning models (Long et al., 2019). However, the system's performance depends on the diversity of the speech data. The model performs poorly on a different speaker with different speech characteristics that were not initially trained on. Personalization of the speaker helps fill that gap by making the ASR model work with the unique characteristics of the individual speaker. Training an ASR model requires high-quality speech data (Long et al., 2019; Radford et al., 2022). Collecting and training user-specific speech data for the model is challenging due to factors that consider user privacy. Most ASR applications are used on lightweight handheld devices with limited processing power. Customization of the model by fine-tuning the model based on user data is very difficult and inefficient due to the significant training parameters. Techniques such as residual adapter for fine-tuning (Tomanek et al., 2021), federated learning on devices by adopting the subset of weights (Jia et al., 2022a), and fine-tuning the model's attention and bias independently (Huang et al., 2021). ASR tasks for low-resource languages and Indo-Aryan languages like Nepali differ due to the language's structure and nature (Bal, 2004). Currently, Parameter Efficient fine-tuning (PEFT) is often used for Large Language Models (LLMs) because it lowers the computation power required to tune the model. The PEFT strategies, like LoRA and DoRA, are used to adapt the models with large trainable parameters. Due to the low use of resources by the adapted and merged model, it allows the large model to be inference using a consumer-grade GPU (Hu et al., 2021). Approaches like LoRA and DoRA are implemented for improving the efficiency of some state-of-the-art ASR models (Joseph and Baby, 2024; Yang et al., 2023). Here, an efficient speaker personalization approach using PEFT, two approaches, LoRA and DoRA, is proposed. The approach uses the low-rank approximation to efficiently adopt the ASR model for the targeted speaker with the limited weight addition, reducing the computation and memory restrictions (Hu et al., 2021; Liu et al., 2024).

2 Related Works

Traditional automatic speech recognition (ASR) architectures employed Hidden Markov Models (HMMs) together with Gaussian Mixture Models (GMM-HMM) to efficiently capture temporal dy-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 189–198

namics and phonetic transitions (Rabiner, 1989). However, HMMs suffered in terms of speaker variability, strict temporal assumptions, and scalability limitations (Chakraborty and Talukdar, 2016). The development of hybrid architectures with Deep Neural Network - Hidden Markov Model (HMM-DNN) combinations improved robustness but at the expense of heavy computational resources (Li et al., 2013). Major architectures used for speech recognition are CNN (LeCun et al., 1989), LSTM (Hochreiter and Schmidhuber, 1997), BiLSTM (Schuster and Paliwal, 1997), RNN (Rumelhart et al., 1986), GRU (Chung et al., 2014). Traditional and recent ASR model uses a combination of the above architectures to perform speech transcription tasks in the Nepali ASR Domain (Regmi and Bal, 2021; Ghimire et al., 2024a). The transformer (Vaswani et al., 2017) based models work great for Nepali ASR tasks (Paudel et al., 2023). For speech personalization, different approaches like the Deep Neural Network (DNN) based acoustic modeling (Hinton et al., 2012), (Singular Value Decomposition) SVD based compression scheme (McGraw et al., 2016), user feedback (Mahesh Krishnamoorthy, 2016), controllable speech synthesis approach (Yang et al., 2023), enhancing quantized model (Zhao et al., 2023) were used. Before that, the data sparsity issue was solved by using speaker dependence with condensed vectors, reducing parameters during model adaptation using some regularization approach (Saon et al., 2013; Snyder et al., 2018; Fan et al., 2020; Sarı et al., 2020). The PEFT strategies used with LLMs includes methods like AdaptFormer (Chen et al., 2022), Visual Prompt Tuning (VPT) (Jia et al., 2022b), Low Rank Adaptation (LoRA) (Hu et al., 2021), Weight-Decomposed Low-Rank Adaptation (DoRA) (Liu et al., 2024), and Scaling & Shifting Your Features (SSF) ((Lian et al., 2022), among those the LoRA and the DoRA are implemented for improving the efficiency of the LLMs like GPT-2/3 (Radford et al., 2019), RoBERTa (Liu et al., 2019), and in some state-of-the-art ASR models (Joseph and Baby, 2024; Yang et al., 2023). Currently, the model is biased per speaker by fine-tuning the model's attention based on methods such as implementing a residual adapter for fine-tuning, federated learning on edge devices, and adopting a subset of weights. There is also a way of targeting a speaker for speech synthesis (Gabryś et al., 2022), which uses the cleaned version of speaker data for ASR Models to process (Wang et al., 2019). Speech data is preprocessed to improve speech parameters fed to the ASR model, thus enhancing the performance of the ASR model in noisy backgrounds (Wu et al., 2017). Experiments have also shown that a small volume of disordered speech of an individual's training data can benefit from Personalized ASR (Tobin and Tomanek, 2021). Fine-tuning the base model has also shown an improvement in the performance of the ASR for the Nepali language using active learning (Ghimire et al., 2023) and PEFT (Ghimire et al., 2024b). The Transformer (Vaswani et al., 2017) model has attention mechanism components, which are used by Low-rank adaptation as the Q, K, and V target modules for speech personalization, which can be used to personalize the ASR system without compromising the model's performance (Joseph and Baby, 2024). Many commercial products like Google Home[1], Amazon Alexa[2], Apple Siri[3], Microsoft Copilot[4], and different voice assistants also perform speech personalization to make interaction with users easier (Hoy, 2018).

The use of LoRA and DoRA is increasing in low-resource languages as well, and this motivated us to perform speaker personalization on the Nepali Language to improve the overall interaction with the ASR system.

3 Methodology

3.1 LoRA-based Speaker Personalization

LoRA (Low-Rank Adaptation)(Hu et al., 2021) technique is adopted to fine-tune the pre-trained weight matrices of the pre-trained Whisper (Radford et al., 2022) model as shown in Figure 2. Following LoRA (Hu et al., 2021), fine-tuning of a original pre-trained weight matrix W_0 (where $W_0 \in R^{d\times k}$; d and k are dimension of input feature vector and output feature vector respectively), the fine-tuning is limited by low-rank decomposition: $W' = W_0 + \Delta W = W_0 + BA$ where $B \in R^{d\times r}$ and $A \in R^{r\times k}$ with rank $r \ll \min(d, k)$. The pre-trained weight W_0 remains frozen, and only A and B are trainable, reducing the computational burden. A and B are multiplied by the same input, and their element-wise additions are calculated, and for an input vector $h = W_0x$, the new forward pass is $h = W'x = Wx + BAx$ (Hu et al., 2021).

[1]https://home.google.com
[2]https://alexa.amazon.com
[3]https://www.apple.com/siri/
[4]https://copilot.microsoft.com/

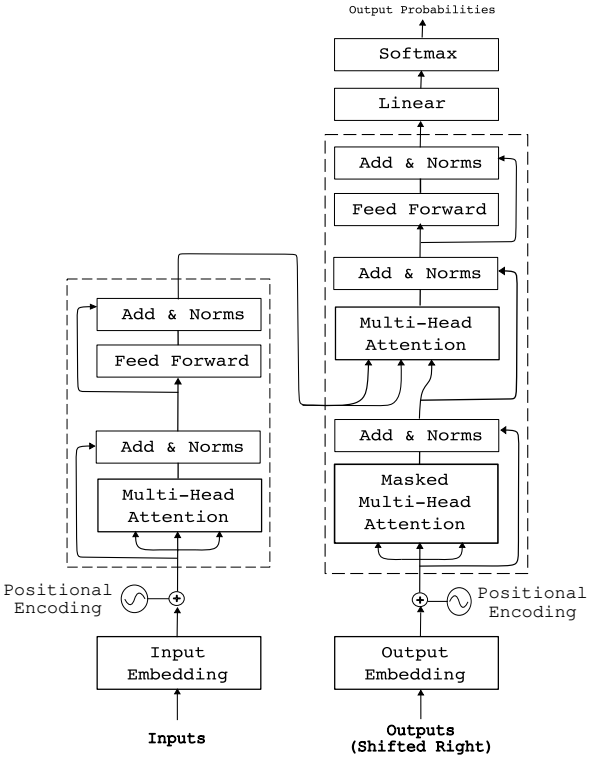

Figure 1: Architecture diagram of the Transformer model

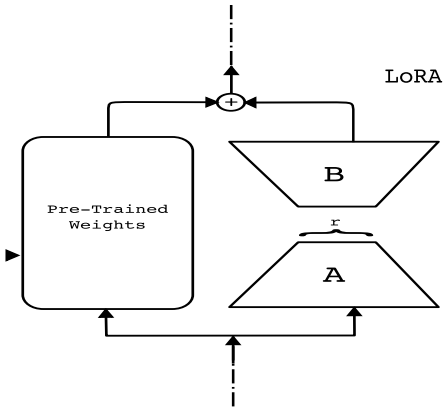

Figure 2: Architecture diagram of LoRA

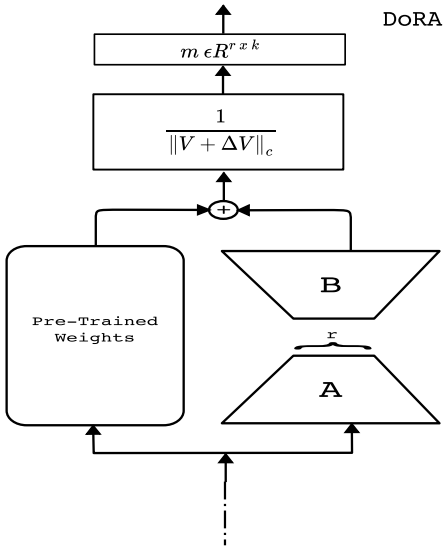

Figure 3: Architecture diagram of DoRA

3.2 DoRA-based Speaker Personalization

DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu et al., 2024) technique is adopted where each weight matrix is mapped in each layer by incorporating another set of weight matrices, as outlined in Figure 3. As per DoRA (Liu et al., 2024), Weight decomposition is outlined as: $W_0 = m * (V/||V||_c) = ||W||_c * (W/||W||_c)$, where $m \in R^{1\times k}$ is the magnitude vector, $V \in R^{d\times k}$ is the directional matrix, and $||.||_c$ denotes the vector-wise norm across each column where d and k are dimension of input feature vector and output feature vector respectively. $W' = m * (V + \Delta V)/(||V + \Delta V||_c) = m * (W_0 + BA)/(||W_0 + BA||_c)$

Here, ΔV is the incremental directional update learned by the product of two low-rank matrices B and A. The matrices $B \in R^{d\times r}$ and $A \in R^{r\times k}$ are initialized according to DoRA's strategy so that W' is equal to W_0 before fine-tuning. DoRA enables more fine-grained updates across attention heads, improving adaptation efficiency (Liu et al., 2024).

3.3 Evaluation Metrics

In this research, Word Error Rate (WER), Character Error Rate (CER), Relative WER (RWER), and Relative CER (RCER) are used to evaluate the model's performance, rank selection, and target module combination selection in PEFT approaches. We discarded Match Error Rate (MER) from our evaluation as it is rarely used in ASR benchmarks and for script-specific nuances of the Nepali language in Devanagari script, where MER offers limited value over CER and complicates interpretations.

3.3.1 WER and CER

WER evaluates the accuracy of text recognition systems at the word level. CER evaluates similarly based on characters. Both metrics measure the percentage of incorrectly recognized words or characters, considering substitutions, insertions, and deletions.

$$\text{WER}\% = \frac{\text{Substitutions+Insertions+Deletions}}{\text{Total Words in Reference}} \times 100$$

Similarly, CER also evaluates the accuracy as WER, but at the character level.

3.3.2 Relative Metrices (RWER and RCER)

The Relative Word Error Rate (Relative WER) and the Relative Character Error Rate (Relative CER) are evaluative measures that examine the performance differences of a specified system compared to a reference baseline.

$$\text{Relative WER}\% = \frac{\text{WER}_{\text{system}} - \text{WER}_{\text{baseline}}}{\text{WER}_{\text{baseline}}} \times 100\%$$

$$\text{Relative CER}\% = \frac{\text{CER}_{\text{system}} - \text{CER}_{\text{baseline}}}{\text{CER}_{\text{baseline}}} \times 100\%$$

3.4 Methodology Details

Two variations of Low-Rank Adapters, namely, LoRA and DoRA, are proposed for speaker personalization to be used with the fine-tuned Whisper (Radford et al., 2022) a transformer (Vaswani et al., 2017) based model from OpenAI in the Huggingface (Wolf et al., 2019) ecosystem. The Proposed approaches for speaker personalization are hereby called **PEFT-LoRA** for the proposed model with LoRA and **PEFT-DoRA** for the proposed approach with DoRA. Another experiment is conducted to check the minimum amount of speech data required to PEFT fine-tune an ASR Model for the optimum rank found as per Table 2 for a set of speakers from Table 1.

Random Gaussian initialization is used to seed the trainable parameters for A; Zero initialization is used for B.ΔW $= BA$ is set at zero at the beginning of training so that the model can gradually learn to adapt to the Nepali language while retaining the pre-learned knowledge from the pre-trained Whisper model even with limited labeled data, making it a perfect fit for low-resource settings. Whereas DoRA / LoRA weights can be applied to any layer, our experiments focus on their integration into the query (W_q), key (W_k), and value (W_v) matrices of the attention mechanism which is in line with findings from previous research, which demonstrated how parameter-efficient approaches can be successfully used on these components to improve model performance (Radford et al., 2022; Hu et al., 2021; Liu et al., 2024; Huang et al., 2020).

Instead of fine-tuning all the parameters, LoRA and DoRA can be used to introduce low-rank learnable parameters update (ΔW) in attention layers, reducing computation cost while maintaining expressiveness (Radford et al., 2022; Liu et al., 2024; Hu et al., 2021).

4 Dataset

A portion of the CommonVoice17 (Ardila et al., 2020) dataset with slice/split of ne-NP/validated was taken for four speakers, and two speakers audio data was taken from the NepDS (Shishir Paudel and Bal Krishna Bal, 2022) dataset by taking the speaker having a commutative speech duration of more than 4 minutes. The data from both datasets is compiled and merged to form a compiled dataset (CommonVoice17 and ILPRL, 2025). The speaker-specific utterances are ordered reversely based on the number of utterances. First, six speakers were selected based on training data ranging from 4 to 18 minutes. Table 1 presents all of the properties of the speaker from the formed dataset labeled under the speaker column where the suffix CV means the speech data from CommonVoice17 and NEPDS means the speech data from the NepDS dataset along with speaker identification number, duration in minutes, number of utterances, and test-train split data. To identify the minimal amount of speech data for LoRA and DoRA PEFT implementation additional dataset with speech data as described in the Speech Range column of Table 4 is prepared.

Speaker	ID	Gender	Duration	Utterances	Train	Test
SpeakerNEPDS1	NS1	M	11.67	216	194	22
SpeakerNEPDS2	NS2	F	17.45	209	188	21
SpeakerCV1	S1	M	8.34	160	144	16
SpeakerCV2	S2	M	9.41	150	135	15
SpeakerCV3	S3	M	5.35	96	86	10
SpeakerCV4	S4	M	4.18	61	54	7

Table 1: Speaker Dataset for combined dataset

5 Experimental Setup

Transformer (Vaswani et al., 2017) Architecture is used for all our experiments implemented through the Huggingface architecture with PyTorch (Paszke et al., 2019) as the codebase. LoRA and DoRA weights are inserted into the Whisper Model Transformer Architecture for training. Every experiment is conducted on "Intel Data Center GPU Max 1550" GPU (Wu et al., 2024). The FT_{base} is a fine-tuned Whisper (Radford et al., 2022) model on OpenSLR54 (Kjartansson et al., 2018). For all the experiments, the Whisper Tokenizer that uses tiktoken (a byte pair tokenizer wrapper) decodes and encodes the dataset used for PEFT (Xu et al., 2023). The dataset (CommonVoice17 and ILPRL, 2025), model, and adapters from this research are

available in HuggingFace[5], and the code used is made available in GitHub[6].

5.1 Transformer Model Architecture

A basic block diagram of the Whisper Model (Transformer Model) is shown in Figure 1. This architecture contains an encoder-decoder structure where the encoder processes into audio features and generates a corresponding token, and the decoder decodes back the output text from the token predicted; the model leverages a multi-head attention mechanism and feed-forward layers to capture both local and global dependencies in the speech data. **Query (Q)**, **Key (K)**, and **Value (V)** are the metrics for self-attention in the transformer model, which PEFT later uses to adopt the Low-Rank Adaptation using LoRA and DoRA.

Layer Normalization and Residual Connections are also leveraged. The model has 1.55 billion parameters, with a 32-layer encoder and decoder, 16 attention heads, and a 1280-dimensional embedding space. It processes 80-dimensional Mel-spectrogram inputs, generates transcriptions using a vocabulary of 51,865 tokens, and employs rotary positional embeddings for efficient sequence modeling (Radford et al., 2022).

5.2 Experiment Details

Whisper Model (Radford et al., 2022) is fine-tuned with OpenSLR54 (Kjartansson et al., 2018) dataset to form a new fine-tuned base model for our experiment called FT_{Base} because the base model from Whisper (Radford et al., 2022) out-of-the-box performed poorly on the Nepali transcription task. The transformer-based encoder and decoder are modified for the Whisper model. Whisper has an encoder and decoder, each with multi-head self-attention and feed-forward networks. In the encoder, every self-attention layer is made up of weight matrices (W_q, W_k, W_v, and W_o) (Xu et al., 2022), initially 1280 × 1280 to match Whisper's embedding size. LoRA replaces these matrices with two smaller matrices, A and B, where A is of size 1280 × r and B is of size r × 1280. The rank of the matrix is obtained by attaching LoRA and DoRA components for ranks of 1 to 128. The final weight update is computed as $\Delta W = A \times B$, and this is added on top of the pre-trained weights. DoRA is also of a similar strategy but decouples rank from input-output dimensions so that updates can be applied separately per attention head. Since Whisper has 16 attention heads per layer, DoRA can better distribute updates across different model parts. LoRA and DoRA may be applied in the cross-attention layers of the decoder and even feed-forward networks, thus allowing fine-tuning without compromising the knowledge of the base model. The formula defines the number of new parameters introduced by LoRA: $LoRA_{params} = N \times m \times (Params\ of\ A, B) = N \times m \times (2 \times C \times r)$ and $DoRA_{params} = N \times m \times (Params\ of\ A, B, V) = N \times m \times (2 \times C \times r + C)$. Where N = Number of Encoder, m= Number of Matrices using DoRA or LoRA Weights, r = $LoRA/DoRA$ Rank, C = Encoder Cell Size (Hu et al., 2021; Liu et al., 2024) and training parameter for each rank is tabulated on Table 2, 3.

We use Rank (r) selection for the speaker adaptation, and further analyze the Query (Q), Key (K), and Value (V) Projection are used to best project the model performance based on the attention modules of the transformer model.

For Rank (r) selection, the Range of rank (r) values is evaluated to find the ideal value that fits with the speaker. The base Model is denoted by $Base$, fine-tuned models are denoted by FT. In the first case, the model is fine-tuned with the Nepali OpenSLR54 Dataset and is denoted as FT_{base}, which will act as a base model for our relative evaluation. In the second case, only the attention layer is fine-tuned (FT_A) as mentioned in Table 2 for every combination of target modules in the observed scenario (Huang et al., 2021). Based on the work by the authors of LoRA and DoRA, multiple sets of Query, Value, and Key Metrics of the attention layers (Hu et al., 2021; Liu et al., 2024) are selected. The subscript of FT_{Base} has the variant whether the FT_{base} is targeted on $C_{Attention}$ (c_attn), $Query(Q)$, $Key(K)$, and $Value(V)$. Experiments on rank are linked on the First Row of the Table. $FT_{A:kv}$ where key value metrics of the attention layer are being fine-tuned, and $FT_{A:qv}$ has query and value of the attention layer being fine-tuned. The number of Trainable Parameters is listed on the row labeled as TP. The ideal rank is determined by conducting several experiments as described above. For ranks above 64 and the training parameters are more significant than $FT_{A:qv}$; thus, ranks ranging from 1 to 32 are considered.

[5]https://hf.co/kiranpantha

[6]https://github.com/kiranpantha/LT-EDI-SPEECH

	Model	TP	WER%							AVG	RCER%	RWER %
			S1	S2	S3	S4	NS1	NS2	AVG	CER%		
	Base	-	88.34	89.93	89.36	82.07	86.54	87.17	87.24	30.17	-	-
	FT_{base}	1.554B	36.85	43.77	58.47	62.41	54.26	53.54	51.55	14.89	-	-
LoRA	$FT_{A:qv}$ $(r = 32)$	15.73M	38.14	40.38	43.16	52.24	17.93	17.26	34.85	8.25	44.59	32.40
LoRA	FT_{qv} $(r = 1)$	0.49M	41.24	40.38	83.16	50.75	20.69	25.22	43.57	8.84	40.62	15.47
LoRA	FT_{qv} $(r = 2)$	0.98M	42.27	32.69	40.00	56.72	20.69	14.16	34.42	8.22	44.80	33.23
LoRA	FT_{qv} $(r = 4)$	1.97M	36.08	34.62	48.42	49.25	21.38	19.47	34.87	8.69	41.64	32.36
LoRA	FT_{qv} $(r = 8)$	3.93M	37.11	40.38	45.30	47.76	16.55	14.16	33.54	7.52	**49.50***	**34.93***
LoRA	FT_{qv} $(r = 16)$	7.86M	43.30	40.38	43.16	52.24	18.62	19.03	36.12	8.55	42.58	29.93
LoRA	FT_{qv} $(r = 32)$	15.73M	38.14	40.38	43.16	53.73	20.69	16.37	35.41	8.42	43.45	31.31
DoRA	$FT_{A:qv}$ $(r = 32)$	15.97M	38.14	40.38	43.16	52.24	20.69	13.72	34.72	8.24	44.66	32.65
DoRA	FT_{qv} $(r = 1)$	0.737M	34.02	40.38	50.53	50.75	23.45	28.32	37.91	9.71	34.79	26.46
DoRA	FT_{qv} $(r = 2)$	1.228M	41.24	32.69	40.00	53.73	20.69	16.81	34.19	8.41	43.52	33.68
DoRA	FT_{qv} $(r = 4)$	2.211M	32.99	34.62	48.42	52.24	20.00	19.91	34.70	8.60	42.24	32.69
DoRA	FT_{qv} $(r = 8)$	3.932M	36.08	40.38	41.05	47.76	16.07	14.16	32.58	7.44	**50.03***	**36.79***
DoRA	FT_{qv} $(r = 16)$	8.110M	42.27	40.38	43.16	50.75	17.24	17.70	35.25	8.36	43.85	31.62
DoRA	FT_{qv} $(r = 32)$	15.97M	38.14	40.38	43.16	52.24	20.69	17.70	35.39	8.41	43.52	31.35

Table 2: Comparision of CER% and WER% accross different rank for LoRA and DoRA, * = selected row for rank based on highest RWER% and RCER%

	K	Q	V	TP	CER%							RCER%	WER%							RWER%
					S1	S2	S3	S4	NS1	NS2	AVG		S1	S2	S3	S4	NS1	NS2	AVG	
LoRA		✓		1.966M	10.04	11.34	11.26	11.62	4.33	7.93	9.42	36.74	45.36	44.23	45.26	52.24	21.38	22.57	38.51	25.30
LoRA	✓			1.966M	9.64	12.15	10.29	11.89	17.45	6.31	11.29	24.18	44.33	50.00	41.05	52.24	33.79	22.57	40.66	21.13
LoRA			✓	3.932M	9.44	9.31	10.1	10.27	3.84	5.50	8.08	45.74	38.14	38.46	46.32	50.75	20.00	18.58	35.38	31.37
LoRA	✓	✓		1.996M	10.44	9.31	10.68	10.00	4.21	6.88	8.59	42.31	47.42	28.85	44.21	50.75	20.69	26.99	36.48	29.23
LoRA		✓	✓	3.932M	10.64	9.72	9.32	11.08	4.21	3.72	8.12	45.47	42.27	40.38	40.00	49.25	22.07	14.16	34.69	32.71
LoRA	✓		✓	3.932M	7.23	10.53	8.54	11.08	3.71	7.28	8.06	**45.87***	28.87	38.46	42.11	52.24	20.00	22.57	34.04	**33.97***
LoRA	✓	✓	✓	5.898M	10.24	10.12	9.13	11.89	3.59	6.47	8.57	42.44	41.24	38.46	41.05	50.75	20.69	21.68	35.64	30.86
DoRA		✓		2.088M	8.84	11.74	10.68	11.62	4.33	8.01	9.20	38.21	41.24	46.15	42.11	52.24	20.69	23.01	37.57	27.12
DoRA	✓			2.088M	9.04	12.15	10.87	12.97	13.24	6.63	10.82	27.33	40.21	50.00	41.05	56.72	30.34	25.66	40.66	21.13
DoRA			✓	4.176M	9.04	9.72	10.29	11.62	3.71	5.58	8.33	44.06	37.11	42.31	45.26	49.25	19.31	19.03	35.38	31.37
DoRA	✓	✓		2.088M	10.44	9.31	10.49	10.27	3.84	8.90	8.87	40.43	45.36	28.85	44.21	52.24	20.69	31.42	37.13	27.97
DoRA		✓	✓	4.176M	10.84	9.72	8.93	10.54	4.46	3.80	8.05	45.94	42.27	40.38	40.00	47.76	22.07	14.60	34.51	33.06
DoRA	✓		✓	4.176M	6.83	10.93	8.54	10.54	3.71	7.20	7.96	**46.54***	31.96	38.46	42.11	47.76	21.38	22.57	34.04	**33.97***
DoRA	✓	✓	✓	6.264M	10.44	10.12	9.13	11.62	3.59	6.72	8.60	42.24	42.27	38.46	41.05	50.75	20.69	21.68	35.82	30.51

Table 3: Comparison of CER% and WER% across different attention layer configurations for LoRA and DoRA for Rank **(r) = 8**; * = selected row for Target Module combination based on highest RWER% and RCER%

For Query, Key, and Value projections, different target modules like $Query(Q)$, $Key(K)$, and $Value(V)$ are experimented to evaluate to analyze the model's performance. The rank is selected as per the initial rank analysis using the RCER and RWER values of the experiment for the given ranks. Each sub-module is taken for the experiment on each set of Q, K, and V parameters. The first column of the Attention Layer has three subdivisions of K, Q, and V, where ✓ means the model is activated on that group of target attention sub-modules. Moreover, the group of Attention Layer per speaker is tabulated below, where $S1$, $S2$, ..., $NS1$, $NS2$, TP, AVG, and $RCER$ have the same meaning as in the Table 2.

For Optimal Speaker Speech data, different quantities of speech data for a speaker were taken from a low of 1 minute to over 13 minutes, as in table 4. After that, the PEFT with the rank selected from optimal rank selection and query, key, and value are taken to apply the LoRA and the DoRA approaches to each commutative speech duration.

Duration (min)	Train Params	LoRA		DoRA	
		CER (%)	WER (%)	CER (%)	WER (%)
1	3.932M	13.71	49.62	13.74	50.31
2	3.932M	11.54	42.29	11.53	43.01
3	3.932M	11.04	42.50	11.04	42.83
4	3.932M	10.03	39.07	10.08	39.24
5	3.932M	9.70	35.84	9.44	37.37
6	3.932M	8.95	36.72	8.81	36.34
7	3.932M	8.81	35.49	8.86	35.38
8	3.932M	6.94	34.11	7.96	33.88
9	3.932M	4.91	22.83	4.93	22.96
10	3.932M	5.89	25.43	5.60	24.80
11	3.932M	6.21	23.20	6.16	22.94
12	3.932M	6.08	24.48	5.95	23.97

Table 4: Comparison of LoRA and DoRA for CER(%) and WER (%) metrics across cumulative speech duration ranges with Rank ($r = 8$), KV Target Model.

6 Results and Discussion

6.1 Rank (r) selection for proposed approach

Rank **8** is selected as the best good value for the rank (r) for personalization of a model, as the rank had the highest RCER and RWER from LoRA and DoRA approaches from Table 2 for the PEFT done on the ranks from 1 to 32 on QV target modules.

6.2 Query, Key and Value Projections

As per Table 3, the model's performance is excellent when the two-attention layers of **Key and Value** are taken for the selected rank of **r = 8**. The Key Value Pair has better RCER and RWER evaluation metrics than other attention modules. The **KV** combination performs better than the **KQ** combination. It gives better results, identical to the resulting pattern obtained from $FT_{A:kv}$, giving better results than $FT_{A:qv}$ (Huang et al., 2021).

6.3 Speech Data Duration for PEFT

The test results from table 4 reflected that just 1 minute of data yielded comparatively poor CER and WER metrics (around 13% CER). However, as more data was utilized, both these metrics continually improved. At around 3 minutes, CER dropped to around 10.5%, and at 5 minutes, to around 9%. Most importantly, when around 10 minutes of training data was used, CER kept dropping at around 5%, indicating good recognition performance.

7 Conclusion

Personalization of speech for the targeted speaker is quite challenging to fine-tune to fit the speech patterns. Using **CER** metrics implementing **LoRA** and **DoRA**, the finding shows **RCER** increment of **49.50%** and **50.03%** respectively. Similarly, for **WER** metrics implementing **LoRA** and **DoRA** shows **RWER** increment of **34.93%** and **36.79%** respectively. This result indicates that the DoRA approach performs better than the LoRA approach for both metrics. Further, the **K, V** combination of target module is found to be best performing in both LoRA and DoRA approaches using both metrics (RCER and RWER) as per Table 2. The result also shows a reduction of **99.74%** of total training parameters used to train and compute using PEFT compared to full fine-tuning.

Taking a system having CER < 5% to be our desired metrics, the findings suggest that around 10 minutes of speaker-dependent data is sufficient for effective fine-tuning with LoRA and DoRA, and it can serve as a good target for speech adaptation, especially personalized ASR applications in low-resource languages.

8 Limitations

Some limitations of the paper are highlighted in this section. Here, the personalization experiments were conducted on a small set of speakers (six in total, five Male and one Female), which may not sufficiently represent the full spectrum of Nepali language variation. Personalization was done in the Nepali language, so this is unclear how the LoRA/DoRA personalization approach will behave in other low-resource languages. Although showing improvement, it's unclear if the model would hold performance with speakers providing noisy, varied, or out-of-domain speech data.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Bal Krishna Bal. 2004. Structure of Nepali Grammer. In *PAN Localization Working Papers*, pages 332–396.

Chandralika Chakraborty and P.H. Talukdar. 2016. Issues and Limitations of HMM in Speech Processing: A Survey. *International Journal of Computer Applications*, 141:13–17.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. *arXiv preprint*. Version Number: 3.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv preprint*. Version Number: 1.

CommonVoice17 and KU ILPRL. 2025. Merged Dataset of CommonVoice17 and NepDS (ILPRL, KU) (Revision 3bfa307).

Zhiyun Fan, Jie Li, Shiyu Zhou, and Bo Xu. 2020. Speaker-aware speech-transformer. *arXiv preprint*. Version Number: 1.

Adam Gabryś, Goeric Huybrechts, Manuel Sam Ribeiro, Chung-Ming Chien, Julian Roth, Giulia Comini, Roberto Barra-Chicote, Bartek Perz, and Jaime Lorenzo-Trueba. 2022. Voice Filter: Few-shot text-to-speech speaker adaptation using voice conversion as a post-processing module. *arXiv preprint*. Version Number: 1.

Rupak Raj Ghimire, Bal Krishna Bal, and Prakash Poudyal. 2023. Active Learning Approach for Fine-Tuning Pre-Trained ASR Model for a Low-Resourced Language: A Case Study of Nepali. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 82–89, Goa University, Goa, India. NLP Association of India (NLPAI).

Rupak Raj Ghimire, Bal Krishna Bal, and Prakash Poudyal. 2024a. A Comprehensive Study of the Current State-of-the-Art in Nepali Automatic Speech Recognition Systems. *arXiv preprint*. Version Number: 1.

Rupak Raj Ghimire, Prakash Poudyal, and Bal Krishna Bal. 2024b. Improving on the Limitations of the ASR Model in Low-Resourced Environments Using Parameter-Efficient Fine-Tuning.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Gerald Penn, and Sanjeev Khudanpur. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. In *ICASSP*, pages 2012–2016. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Matthew B. Hoy. 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly*, 37(1):81–88.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. Version Number: 2.

Yan Huang, Jinyu Li, Lei He, Wenning Wei, William Gale, and Yifan Gong. 2020. Rapid RNN-T Adaptation Using Personalized Speech Synthesis and Neural Language Generator. In *Interspeech 2020*, pages 1256–1260. ISCA.

Yan Huang, Guoli Ye, Jinyu Li, and Yifan Gong. 2021. Rapid Speaker Adaptation for Conformer Transducer: Attention and Bias Are All You Need. In *Interspeech 2021*, pages 1309–1313. ISCA.

Junteng Jia, Jay Mahadeokar, Weiyi Zheng, Yuan Shangguan, Ozlem Kalinli, and Frank Seide. 2022a. Federated Domain Adaptation for ASR with Full Self-Supervision. In *Interspeech 2022*, pages 536–540. ISCA.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022b. Visual Prompt Tuning. *arXiv preprint*. Version Number: 2.

George Joseph and Arun Baby. 2024. Speaker Personalization for Automatic Speech Recognition using Weight-Decomposed Low-Rank Adaptation. In *Interspeech 2024*, pages 2875–2879. ISCA.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 52–55. ISCA.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551.

Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. 2013. Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 312–317.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. 2022. Scaling & Shifting Your Features: A New Baseline for Efficient Model Tuning. *arXiv preprint*. Version Number: 3.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-Decomposed Low-Rank Adaptation. *arXiv preprint*. Version Number: 6.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. Version Number: 1.

Yanhua Long, Yijie Li, Shuang Wei, Qiaozheng Zhang, and Chunxia Yang. 2019. Large-Scale Semi-Supervised Training in Deep Learning Acoustic Model for ASR. *IEEE Access*, 7:133615–133627.

Matthias Paulik Mahesh Krishnamoorthy. 2016. Improving Automatic Speech Recognition Based on User Feedback.

Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Hasim Sak, Alexander Gruenstein, Francoise Beaufays, and Carolina Parada. 2016. Personalized Speech recognition on mobile devices. *arXiv preprint*. Version Number: 2.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv preprint*. Version Number: 1.

Shishir Paudel, Bal Krishna Bal, and Dhiraj Shrestha. 2023. Large Vocabulary Continous Speech Recognition for Nepali Language using CNN and Transformer. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 328–333, Vienna, Austria. NOVA CLUNL, Portugal.

L.R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint*. Version Number: 1.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Sunil Regmi and Bal Krishna Bal. 2021. An End-to-End Speech Recognition for the Nepali Language. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 180–185, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA.

George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny. 2013. Speaker adaptation of neural network acoustic models using i-vectors. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 55–59, Olomouc, Czech Republic. IEEE.

Leda Sarı, Niko Moritz, Takaaki Hori, and Jonathan Le Roux. 2020. Unsupervised Speaker Adaptation using Attention-based Speaker Memory for End-to-End ASR. *arXiv preprint*. Version Number: 1.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Shishir Paudel and Bal Krishna Bal. 2022. NepDS Dataset - Information and Language Processing Research Lab (ILPRL).

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, Calgary, AB. IEEE.

Jimmy Tobin and Katrin Tomanek. 2021. Personalized Automatic Speech Recognition Trained on Small Disordered Speech Datasets. *arXiv preprint*. Version Number: 1.

Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. 2021. Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech. *arXiv preprint*. Version Number: 1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint*. Version Number: 7.

Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno. 2019. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. In *Interspeech 2019*, pages 2728–2732. ISCA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*. Version Number: 5.

Bo Wu, Kehuang Li, Fengpei Ge, Zhen Huang, Minglei Yang, Sabato Marco Siniscalchi, and Chin-Hui Lee. 2017. An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1289–1300.

Hui Wu, Yi Gan, Feng Yuan, Jing Ma, Wei Zhu, Yutao Xu, Hong Zhu, Yuhua Zhu, Xiaoli Liu, Jinghui Gu, and Peng Zhao. 2024. Efficient LLM inference solution on Intel GPU. *arXiv preprint*. Version Number: 2.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *arXiv preprint*. Version Number: 1.

Sheng Xu, Yanjing Li, Teli Ma, Bohan Zeng, Baochang Zhang, Peng Gao, and Jinhu Lv. 2022. TerViT: An Efficient Ternary Vision Transformer. *arXiv preprint*. ArXiv:2201.08050 [cs].

Karren Yang, Ting-Yao Hu, Jen-Hao Rick Chang, Hema Swetha Koppula, and Oncel Tuzel. 2023. Text is All You Need: Personalizing ASR Models using Controllable Speech Synthesis. *arXiv preprint*. ArXiv:2303.14885 [eess].

Qiuming Zhao, Guangzhi Sun, Chao Zhang, Mingxing Xu, and Thomas Fang Zheng. 2023. Enhancing Quantised End-to-End ASR Models via Personalisation. *arXiv preprint*. ArXiv:2309.09136 [cs].

An Overview of the Misogyny Meme Detection Shared Task for Chinese Social Media

Bharathi Raja Chakravarthi[1], **Rahul Ponnusamy**[2], **Ping Du**[1], **Xiaojian Zhuang**[1], **Saranya Rajiakodi**[3], **Paul Buitelaar**[2], **Premjith B**[4], **Bhuvaneswari Sivagnanam**[3] **Anshid Kizhakkeparambil**[5], **Lavanya S.K.**[6]

[1] School of Computer Science, University of Galway, Ireland
[2] Data Science Institute, University of Galway, Ireland
[3] Department of Computer Science, Central University of Tamil Nadu, India
[4] Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[5] Mar Thoma College, Chungathara, Kerala, India
[6] Madras Institute of Technology Campus, Anna University, India

Abstract

The increasing prevalence of misogynistic content in online memes has raised concerns about their impact on digital discourse. The culture-specific images and informal usage of text in the memes present considerable challenges for the automatic detection systems, especially in low-resource languages. While previous shared tasks have addressed misogyny detection in English and several European languages, misogynistic meme detection in the Chinese has remained largely unexplored. To address this gap, we introduced a shared task focused on binary classification of Chinese language memes as misogynistic or non-misogynistic. The task featured memes collected from the Chinese social media and annotated by native speakers. A total of 45 teams registered, with 8 teams submitting predictions from their multimodal models integrating textual and visual features through diverse fusion strategies. The best-performing system achieved a macro F1-score of 0.93035, highlighting the effectiveness of lightweight pretrained encoder fusion. This system used the Chinese BERT and DenseNet-121 for text and image feature extraction, respectively. A feedforward network was trained as a classifier using the features obtained by concatenating text and image features.

1 Introduction

The proliferation of social media and other online platforms provides users with powerful means to share visual (image and video), audio, and textual content as a form of expressing opinions and beliefs (Huang et al., 2019). Memes are typically used for expressing one's opinion through humorous or satirical images overlaid with text (Zhong et al., 2022), (Aslam et al., 2022), (Aggarwal et al., 2021). However, the dual-mode characteristic, which contains both text and images, and the speed at which the memes are spread make them a powerful tool for communication, which in turn makes them carriers of offensive and hate content, including misogyny (Jindal et al., 2024), (Chakravarthi et al., 2025). Furthermore, misogynistic memes frequently trivialize the violence against women and discriminatory attitudes by employing humor and irony (Kumari, 2021).

Misogynistic memes are difficult to detect due to the multimodal ambiguity (Hakimov et al., 2022), (Rizzi et al., 2024), (Kiela et al., 2020). Moreover, the models that are used for detecting the misogynistic memes should be capable of capturing the cultural nuances, sarcasm, and symbolism, mainly because of the diverse linguistic and social settings (Kumari et al., 2024). Much of the prior research has focused on English and European languages (Fersini et al., 2022a). For example, the MAMI shared task at SemEval-2022 targeted misogyny detection in English memes using image-text combinations, while SemEval-2024 addressed multimodal persuasion techniques in memes across multiple languages, including English and Bulgarian. Further, recent shared tasks such as LT-EDI@EACL 2024 and DravidianLangTech@NAACL 2025 have explored this problem in Tamil and Malayalam (Chakravarthi et al., 2024, 2025). Despite these advancements, misogynistic meme detection in Chinese—one of the most widely spoken languages globally (Julian, 2020)—has remained largely unexplored.

To address this gap, we introduced the shared task on misogyny meme detection in the Chinese. This task focuses on the binary classification of memes sourced from real-world Chinese social media platforms, annotated by native speakers as ei-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 199–207

ther misogynistic or non-misogynistic. The memes reflect various forms of gender-based prejudice, often expressed through culturally grounded visual cues and colloquial language.

This shared task encouraged participants to develop multimodal models capable of jointly interpreting visual and textual cues. The multilingual complexity of Chinese, the presence of colloquial and coded language, and the use of culturally specific imagery added rich layers of difficulty. By bringing together researchers across NLP, computer vision, and digital humanities, the task aims to foster tools and insights that are socially aware, linguistically grounded, and technically robust. Moreover, code-mixing between Chinese and English is common on Chinese social media platforms (Zhang, 2012). The code-mixing involves vocabulary substitution, where English words, abbreviations, brand names, or proprietary terms are inserted into Chinese sentences (Guo, 2023). This creates a lot of challenges for building AI models for code-mixed Chinese social media data. The language identification and segmentation is one of the major challenges for models, particularly embedding models. The unavailability of the high-quality, large-scale corpus is another challenge for Chinese-English code-mixed data modeling. Pretrained language models, even large multilingual ones, often perform poorly on code-mixed inputs compared to monolingual or standard multilingual data (Kodali et al., 2024).

A total of 45 teams registered for the shared task, of which 8 teams successfully submitted their system outputs for final evaluation. The submitted systems employed a range of multimodal strategies, combining textual and visual features through various fusion techniques. The CUET_320 secured the first position by achieving the highest macro F1-score of 0.93035. This demonstrates the effectiveness of lightweight fusion of pretrained text and image encoders. Teams CUET_Ignite (0.91775) and Team_Luminaries_0227 (0.90355) attained the second and third positions. These teams showed that transformer-based language models and vision encoders can perform effectively when optimized with minimal compute. Despite these successes, ongoing challenges include class imbalance, noisy visual-text alignment, and the handling of code-mixed linguistic patterns in Chinese memes.

2 Related Work

2.1 Misogyny Meme Detection

Detecting misogynous memes is a challenging task due to the multimodal ambiguity. Researchers have proposed several approaches and models to address this challenge. In (Gu et al., 2022), the authors proposed a multi-modal, multi-task variational autoencoder designed to detect misogynous memes. This model learns a co-representation from both images and text together in a latent feature space to find misogynous memes and classify them into specific categories. In (Agrawal and Mamidi, 2022), the authors proposed visual linguistic models with transfer learning. These models followed a task-specific pretraining of the models. The authors (Mahesh et al., 2024) proposed an approach by combining models pretrained for text and image data in Tamil and Malayalam. The authors conducted experiments using BERT+ResNet-50, MuRIL+ResNet-50, and mBERT+ResNet-50. Among these models, mBERT+ResNet-50 and MuRIL+ResNet-50 achieved the highest macro F1-scores of 0.73 and 0.87 for Tamil and Malayalam data, respectively.

The authors (Roy et al., 2024) designed a Bidirectional Long Short-Term Memory (Bi-LSTM) with BERT embeddings network for detecting misogynistic comments in Bengali along with a comprehensive dataset. The study proposed by (Chinivar et al., 2024) used the XLM-R model along with two image transformer models, ViT and Swin, on a benchmarked meme dataset. Here, the authors concatenated the embeddings generated from text and image data and fed them into a single-layered neural network classifier for classification. In (Gitanjali et al., 2024), the authors proposed a Multimodal Multi-hop Chain-of-Thought (M3Hop-CoT) framework for misogynous meme identification. They integrated entity-object-relationship information with a multimodal CoT module, thereby providing emotional cues, target awareness, and contextual knowledge for meme analysis. Finally, a CLIP-based classifier was employed for classification.

In (Farinango Cuervo and Parde, 2022), the authors used contrastive learning for the multimodal detection of misogynistic memes. Contrastive learning ensured that memes labeled as misogynistic were grouped together by differentiating them from non-misogynistic ones. Detecting misogynistic memes in the Chinese language is a complex and emerging field. CHMEMES (Gu et al., 2024)

is a dataset for detecting harmful memes in the Chinese language. Moreover, the authors proposed a multimodal framework, semantic contrastive alignment framework (SCARE) to implement the task. The proposed framework learns both cross-modal and intra-modal information, with a cross-modal contrast alignment objective and an intra-modal contrast objective.

2.2 Related Shared tasks

Research on misogyny and hate speech detection has been significantly shaped by early shared tasks. Evalita 2018 and IberEval 2018 introduced the Automatic Misogyny Identification (AMI) challenge for English and Italian, focusing on text-based detection of gendered hate speech (Fersini et al., 2018a,b). These were among the first systematic efforts in the domain. Later, SemEval 2019 Task 5 expanded this scope to multilingual hate speech, targeting both misogyny and immigrant hate in English and Spanish. It introduced subtasks on aggression level classification, establishing the need for nuanced annotation in hate speech datasets (Basile et al., 2019).

Multimodal shared tasks further pushed boundaries by incorporating image and text jointly. The Memotion Analysis tasks at SemEval 2020 addressed meme-based sentiment and emotion classification in English, providing large annotated meme datasets and benchmarks, enabling exploration of the affective dimensions of memes (Sharma et al., 2020; Patwa et al., 2022).

The MAMI (Multimedia Automatic Misogyny Identification) task at SemEval-2022 focused explicitly on misogyny detection in English memes, with subtasks for binary and fine-grained classification of misogynistic intent using both image and text (Fersini et al., 2022b). The recent SemEval-2024 Task targeted persuasion technique detection in memes using English training data, with test sets in English, Bulgarian, North Macedonian, and Arabic, and included both text-only and multimodal subtasks (Dimitrov et al., 2024).

In line with these developments, the LT-EDI@EACL 2024 shared task introduced misogynistic meme detection in Tamil and Malayalam as part of a multitask classification challenge (Chakravarthi et al., 2024). Building on this, the DravidianLangTech@NAACL 2025 shared task extended the focus to dedicated misogyny meme detection (Chakravarthi et al., 2025). Our current shared task, organized at LT-EDI@LDK 2025, continues this progression by targeting Chinese-language misogynistic memes. This task aims to broaden multimodal hate speech research to new linguistic and cultural domains and addresses the scarcity of annotated resources in Chinese for multimodal misogyny detection.

3 Task Description

The shared task on Misogyny Meme Detection is organized as part of the LT-EDI workshop at LDK 2025[1]. This task focuses on a multimodal classification problem that targets the automatic detection of misogynistic content in memes.

Memes present a unique challenge for automated hate speech detection due to their multimodal nature, combining both images and text to convey meaning. This shared task aims to address the rising concern of misogynistic content, particularly in Chinese online social media platforms. Couple of examples from the dataset can be seen in the Figure 1. The task emphasizes the need for robust computational models that can effectively integrate textual and visual information to identify misogyny. Participants are required to develop systems that classify memes into two categories:

- Misogyny
- Not-Misogyny

This task supports ongoing research in multimodal hate speech detection and contributes to the development of ethical AI tools capable of identifying harmful content online.

3.1 Dataset and Evaluation

Dataset: The dataset utilized in this study comprises memes collected from various Chinese social media platforms. Each meme consists of two modalities: a meme image and its corresponding textual content, which is a transcription contained within the meme. The annotation provided are at the level of the entire meme. These annotations are binary categorizing each meme as either containing misogynistic content or not.

Task Phases: The task is organized into two distinct phases:

- **Training and Development Phase:** In this initial phase, participants are provided with a labeled datset. This dataset is intended for

[1] https://codalab.lisn.upsaclay.fr/competitions/21880

Figure 1: Example memes from the dataset with the translation

the training and validation of their models. It contains a collection of both misogynistic and non-misogynistic memes, each accompanied by its respective image and textual data.

- **Test Phase:** Following the training and development phase, a separate, unlabeled test set will be distributed. Participants are required to submit their system's predictions for this test set. The final evaluation of the submitted systems will be performed using this unseen data.

Evaluation Metric: The primary metric for evaluating performance in this task is the macro F1-score. This metric is chosen because it considers the performance across both the misogynistic and non-misogynistic classes equally. The macro F1-score is particularly well-suited for classification tasks where there may be an imbalance in the number of samples per class, thereby ensuring a fair and comprehensive assessment of model performance on both categories.

4 Methodology of Participants

This competition included eight entrants, each presenting a distinct approach to misogynist meme identification. Most teams employed various preprocessing methods, data augmentation strategies, and model architectures to build multimodal pipelines that integrated textual and visual features. The approaches ranged from deep learning and transformer-based models to traditional machine learning, with different fusion techniques applied to effectively combine modalities. A summary of each team's core methodology is provided in the following section.

Team_Luminaries_0227 (Adnan Faisal, 2025) used a multimodal fusion strategy to tackle the Misogyny Meme Detection task. After testing with XLM-RoBERTa (Li et al., 2021) and DistilBERT (Sanh et al., 2019), they used the Chinese BERT model (bert-base-chinese) for textual analysis. They evaluated VGG16, ViT, and ResNet in the visual domain before deciding on VGG16 because to its effectiveness. The imagess were preprocessed to 224x224 pixels and enhanced with rotations and random flips. Jieba (Zhang et al., 2019) was used to clean and tokenize the textual data. In order to merge text and image information and feed them into a classifier for the ultimate prediction, the researchers used early fusion. The Adam optimizer was used for training, with a batch size of 32, a learning rate of 2e-5, and early stop-

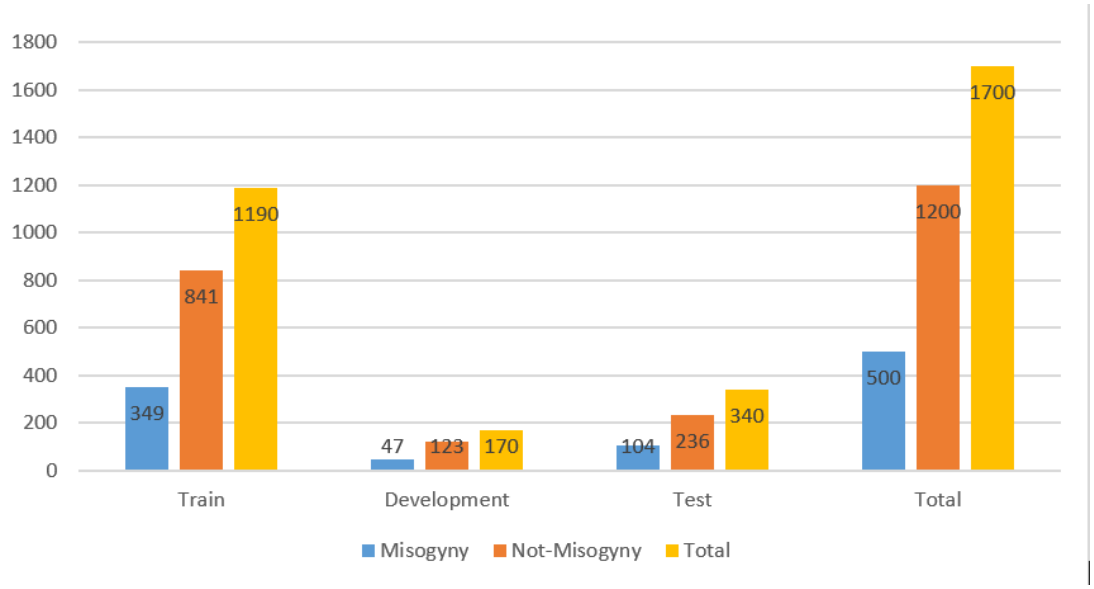

Figure 2: Dataset Statistics

ping determined by validation loss. Weighted loss functions were used to remedy class imbalance. On the Malayalam dataset, their method produced a macro F1-score of 0.87, and on the Tamil dataset, 0.73.

CUET_12033 (Mehreen Rahman, 2025) proposed a compact yet effective multimodal pipeline for detecting misogyny in Chinese memes. They preprocessed text by cleaning, converting to Simplified Chinese, tokenizing with Jieba, and transliterating using pypinyin. Images were resized to 224×224 RGB and enhanced for brightness and contrast. To address class imbalance, misogynistic samples were augmented with brightness adjustment, grayscale conversion, and posterization. For unimodal models, CharBERT-base-Chinese was enhanced with a BiLSTM, while image features were extracted using CLIP, ViT, ResNet-50, and EfficientNet-B0. Their best multimodal model fused CharBERT+BiLSTM (Ma et al., 2020) and ViT embeddings using a gated multimodal unit (GMU), followed by a classifier. Training included class-weighted loss, mixed precision, learning rate scheduling, dropout, early stopping, and gradient clipping for stability.

CUET_Ignite (MD.Mahadi Rahman, 2025) proposed a robust multimodal pipeline for misogyny detection in Chinese memes. Their preprocessing involved cleaning and transliterating text (using Jieba and pypinyin) and enhancing 224×224 RGB images. To address class imbalance, they applied augmentation only on misogynistic samples—using image transformations and back-translation. For modeling, they used CharBERT with or without BiLSTM for text and explored CLIP, ViT, ResNet-50, and EfficientNet-B0 for image features. Text and image embeddings were projected into a joint 512-dim space and fused via multi-head attention, followed by a two-layer MLP. Training used Adam/AdamW with class-weighted loss and optimization strategies like mixed precision, scheduling, and early stopping.

SSNCSE (Sreeja K, 2025) developed a multimodal classification system using frozen pretrained models for efficient misogyny detection in Chinese memes. Text was cleaned and tokenized using BERT-base, while images were resized and normalized before extracting 2048-dim features with ResNet-50. The 768-dim text and 2048-dim image embeddings were concatenated into a 2816-dim vector and passed through a lightweight feed-forward classifier with ReLU and dropout. Only classifier layers were trained using Adam (LR=1e-4, batch size=8) over 5 epochs, optimizing cross-entropy loss. Final predictions were stored in CSV format.

CUET_320 (Madiha Ahmed Chowdhury, 2025) developed an efficient multimodal system for misogyny meme detection by preprocessing text with regex cleaning, Jieba tokenization, stopword removal, and truncation, while images were resized, histogram-enhanced, and normalized. To balance classes, misogynistic samples were augmented using brightness, grayscale, and posteriza-

tion. Features were extracted via Chinese BERT, XLM-R, mBERT (text), and CNNs like ResNet-50, DenseNet-121, Inception V3 (image). Their best model combined Chinese BERT and DenseNet-121 features through concatenation and classified them using an Adam-optimized feedforward network, showing the effectiveness of simple fusion.

CVF_NITT (Radhika K T, 2025) employed Visual-Language Models (VLMs), particularly CLIP, to detect sexist elements in Chinese memes by aligning OCR-extracted text and image features in a shared 512-dimensional space. They used a lightweight logistic regression classifier on these embeddings (CLIP+LR) and compared it against traditional fusion setups like BERT+ResNet-50 and CNN+Inception V3 followed by MLPs. Their early-fusion CLIP-based approach proved more effective for cross-modal integration.

CUET's_White_Walkers (Md Mubasshir Naib, 2025) created a modular multimodal system for sexism meme identification by preprocessing text (cleaned, lowercased, tokenized to 128 tokens) and images (resized to 224×224, normalized). Images were enhanced through flipping and cropping. They used transformer-based models (mBERT, MuRIL, BERT-base-Chinese), deep (GloVe/Keras embeddings with CNN, BiLSTM-CNN, CNN-GRU), and classical (TF-IDF, BoW with ML classifiers) to extract features from images and pretrained CNNs (ResNet50, DenseNet121, InceptionV3) for text. Using an early fusion of BERT-base-Chinese and ResNet50, their optimal multimodal setup outperformed all unimodal baselines with an F1 score of 0.8541.

5 Results and Discussion

All the participant's system were evaluated using a macro F1-score. Among the participating teams, CUET_320 achieved the highest performance with a macro F1-score of 0.93035 which secured first rank with Chinese BERT and DenseNet-121 combination . This was followed by CUET_Ignite which obtained a macro F1-score of 0.91775 and ranked second. The third position was held by Team_Luminaries_0227 which attained a macro F1-score of 0.90355 with Chinese BERT and VGG16 combination. These results indicate strong overall performance and highlight the effectiveness of the multimodal approaches used by the top teams. Most teams froze the pretrained encoders (e.g., ResNet50, BERT, CLIP) and trained only the classification layers, which helped reduce overfitting and training cost. Fusion techniques such as cross-attention and multi-head attention proved to be highly effective in modeling the relationship between images and text. Several teams addressed class imbalance using weighted loss functions or data augmentation, which improved minority class recall. Teams used optimization techniques such as FP16 training and batch accumulation to efficiently utilize limited compute resources (e.g., Kaggle P100 GPUs). These methods significantly improved the misogyny detection in low-resource high-noise meme datasets.

6 Conclusion

In this shared task on misogynistic meme detection, we challenged participants to build a multimodal classification systems capable of identifying misogyny in memes by analyzing both textual and visual components. The participating teams explored a variety of deep learning architectures, including combinations of pretrained CNNs, transformer-based text encoders, and fusion mechanisms. The top performing systems demonstrated the importance of careful preprocessing and enhancement, effective multimodal fusion, and efficient training pipelines to address imbalance and limited resources. These findings confirm that tackling harmful online content like misogynistic memes requires context-aware, cross-modal learning approaches.

Future work may explore zero-shot detection using large vision-language models (e.g., BLIP-2, LLaVA), prompt-tuning for meme analysis, and generating explanations for detected misogyny to aid in interpretability and platform moderation.

Acknowledgments

The authors, Bharathi Raja Chakravarthi and Paul Buitelaar was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight) and author Rahul Ponnusamy was supported by the Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Momtazul Arefin Labib Hasan Murad Adnan Faisal, Shiti Chowdhury. 2025. Team_luminaries_0227@lt-edi-2025: A transformer-based fusion approach to

Team Name	Run	F1-score	rank
CUET_320_run1 (Madiha Ahmed Chowdhury, 2025)	**1**	**0.93035**	**1**
CUET_Ignite_run2 (MD.Mahadi Rahman, 2025)	2	0.91775	2
Team_Luminaries_0227 (Adnan Faisal, 2025)	3	0.90355	3
CUET's_White_Walkers (Md Mubasshir Naib, 2025)	2	0.85421	4
CVF_NITT (Radhika K T, 2025)	-	0.73622	5
CUET_12033_run1 (Mehreen Rahman, 2025)	1	0.70898	6
SSNCSE_Sreeja K (Sreeja K, 2025)	1	0.70345	7
CUET_Fog_run1	1	0.49514	8

Table 1: list of the teams participated in the shared task

misogyny detection in chinese memes. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Apeksha Aggarwal, Vibhav Sharma, Anshul Trivedi, Mayank Yadav, Chirag Agrawal, Dilbag Singh, Vipul Mishra, and Hassène Gritli. 2021. Two-way feature extraction using sequential and multimodal approach for hateful meme classification. *Complexity*, 2021(1):5510253.

Samyak Agrawal and Radhika Mamidi. 2022. LastResort at SemEval-2022 task 5: Towards misogyny identification using visual linguistic model ensembles and task-specific pretraining. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 575–580, Seattle, United States. Association for Computational Linguistics.

Nida Aslam, Irfan Ullah Khan, Teaf I Albahussain, Nouf F Almousa, Mizna O Alolayan, Sara A Almousa, and Modhi E Alwhebi. 2022. Medeep: A deep learning based model for memotion analysis. *Mathematical Modelling of Engineering Problems*, 9(2).

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid K A. 2025. Findings of the shared task on misogyny meme detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 721–731, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Sneha Chinivar, MS Roopa, JS Arunalatha, and KR Venugopal. 2024. Identification of misogynistic memes using transformer models. In *International Conference on Advanced Communications and Machine Intelligence*, pages 107–116. Springer.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Charic Farinango Cuervo and Natalie Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 785–792, Seattle, United States. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022a. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022b. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*,

pages 533–549, Seattle, United States. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, Nicole Novielli, and Viviana Patti, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian : Proceedings of the Final Workshop 12-13 December 2018, Naples*, Collana dell'Associazione Italiana di Linguistica Computazionale, pages 59–66. Accademia University Press, Torino. Code: EVALITA Evaluation of NLP and Speech Tools for Italian : Proceedings of the Final Workshop 12-13 December 2018, Naples.

Elisabetta Fersini, Paolo Rosso, Maria Anzovino, and 1 others. 2018b. Overview of the task on automatic misogyny identification at IberEval 2018. *Ibereval@ sepln*, 2150:214–228.

Kumari Gitanjali, kirtan Jain, and Asif Ekbal. 2024. M3hop-cot: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 22105–22138.

Tianlong Gu, Mingfeng Feng, Xuan Feng, and Xuemin Wang. 2024. Scare: A novel framework to enhance chinese harmful memes detection. *IEEE Transactions on Affective Computing*.

Yimeng Gu, Ignacio Castro, and Gareth Tyson. 2022. MMVAE at semeval-2022 task 5: A multi-modal multi-task VAE on misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 700–710.

Yuanhao Guo. 2023. The Code-Mixing of Chinese and English Among Chinese College Students: A Qualitative Study. In *2022 4th International Conference on Literature, Art and Human Development (ICLAHD 2022)*, pages 73–89. Atlantis Press.

Sherzod Hakimov, Gullal Singh Cheema, and Ralph Ewerth. 2022. TIB-VA at SemEval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 756–760, Seattle, United States. Association for Computational Linguistics.

Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167:26–37.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. MISTRA: Misogyny Detection through Text–Image Fusion and Representation Analysis. *Natural Language Processing Journal*, 7:100073.

George Julian. 2020. What are the most spoken languages in the world. *Retrieved May*, 31(2020):38.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Prashant Kodali, Anmol Goel, Likhith Asapu, Vamshi Krishna Bonagiri, Anirudh Govil, Monojit Choudhury, Ponnurangam Kumaraguru, and Manish Shrivastava. 2024. From human judgements to predictive models: Unravelling acceptability in code-mixed sentences. *arXiv preprint arXiv:2405.05572*.

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3Hop-CoT: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22105–22138, Miami, Florida, USA. Association for Computational Linguistics.

Sapna Kumari. 2021. Meme culture and social media as gendered spaces of dissent and dominance. *Journal of Visual Literacy*, 40(3-4):215–232.

Bing Li, Yujie He, and Wenjin Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Charbert: Character-aware pre-trained language model. *arXiv preprint arXiv:2011.01513*.

MD. SHAFIQUL HASAN Ashim Dey Madiha Ahmed Chowdhury, Lamia Tasnim Khan. 2025. Cuet_320@lt-edi-2025: A multimodal approach for misogyny meme detection in chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Sidharth Mahesh, D Sonith, Gauthamraj Gauthamraj, G Kavya, Asha Hegde, and H Shashirekha. 2024. MUCS@ LT-EDI-2024: Exploring Joint Representation for Memes Classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287.

Jidan Al Abrar Md Mehedi Hasan Md Siddikul Imam Kawser Mohammad Shamsul Arefin Md Mubasshir Naib, Md Mizanur Rahman. 2025. Cuets_white_walkers@lt-edi-2025: A multimodal framework for the detection of misogynistic memes in chinese online content. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Mohammad Oman Mohammad Shamsul Arefin MD.Mahadi Rahman, Mohammad Minhaj Uddin. 2025. Cuet_ignite@lt-edi-2025: A multimodal transformer-based approach for detecting misogynistic memes in chinese social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Nabilah Tabassum Samia Rahman Hasan Murad Mehreen Rahman, Faozia Fariha. 2025. Cuet_12033lt-edi-2025: Misogyny detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Parth Patwa, Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Findings of memotion 2: Sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Sitara K Radhika K T. 2025. Cvf-nitt@lt-edi-2025:misogynydetection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Giulia Rizzi, David Gimeno-Gómez, Elisabetta Fersini, and Carlos-D Martínez-Hinarejos. 2024. Pink at exist2024: a cross-lingual and multi-modal transformer approach for sexism detection in memes. *Working Notes of CLEF*.

Debopriya Deb Roy, Israt Moyeen Noumi, and Md Aminur Rahman. 2024. Towards Better Misogyny Detection in Bangla: Improved Dataset and Cutting-Edge Model Evaluation. In *2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pages 1–6. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Bharathi B Sreeja K. 2025. Ssncse@lt-edi-2025:detecting misogyny memes using pretrained deep learning models. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Wei Zhang. 2012. Chinese-English code-mixing among China's netizens: Chinese-English mixed-code communication is gaining popularity on the Internet. *English Today*, 28(3):40–52.

Xianwei Zhang, Peng Wu, Jiuming Cai, and Kun Wang. 2019. A Contrastive Study of Chinese Text Segmentation Tools in Marketing Notification Texts. *Journal of Physics: Conference Series*, 1302(2):022010.

Qi Zhong, Qian Wang, and Ji Liu. 2022. Combining knowledge and multi-modal fusion for meme classification. In *International Conference on Multimedia Modeling*, pages 599–611. Springer.

Findings of the Shared Task Multilingual Bias and Propaganda Annotation in Political Discourse

Shunmuga Priya Muthusamy Chinnan[1], Bharathi Raja Chakravarthi[1], Meghann L. Drury-Grogan[3], Senthil Kumar B[4], Saranya Rajiakodi[5], Angel Deborah Suseelan[6], Jason Joachim Carvalho[1]

[1] School of Computer Science, University of Galway, Ireland
[3] Department of Enterprise and Technology, Atlantic Technological University, Ireland
[4] Department of Information Technology, Velammal Institute of Technology, India
[5] Department of Computer Science, Central University of Tamil Nadu, India
[6] Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, India

Abstract

The Multilingual Bias and Propaganda Annotation task focuses on annotating biased and propagandist content in political discourse across English and Tamil. This paper presents the findings of the shared task on bias and propaganda annotation task. This task involves two sub tasks, one in English and another in Tamil, both of which are annotation task where a text comment is to be labeled. With a particular emphasis on polarizing policy debates such as the US Gender Policy and India's Three Language Policy, this shared task invites participants to build annotation systems capable of labeling textual bias and propaganda. The dataset was curated by collecting comments from YouTube videos. Our curated dataset consists of 13,010 English sentences on US Gender Policy, Russia-Ukraine War and 5,880 Tamil sentences on Three Language Policy. Participants were instructed to annotate following the guidelines at sentence level with the bias labels that are fine-grained, domain specific and 4 propaganda labels. Participants were encouraged to leverage existing tools or develop novel approaches to perform fine-grained annotations that capture the complex socio-political nuances present in the data.

1 Introduction

Social media platforms have become important medium for communication, enabling the widespread exchange and access to information from diverse sources (Datta et al., 2021). However, this open ecosystem is increasingly filled with harmful content, including various forms of misinformation such as propaganda, conspiracy theories, and biased narratives (Zubiaga et al., 2016). The rapid scale and sophistication of such content demand solutions beyond manual fact checking (Nakov and Da San Martino, 2020). Consequently, developing automated methods to detect and mitigate biased and propagandist content has become an urgent research priority (Zaghouani et al., 2024; Aksenov et al., 2021).

The term bias is defined as "Bias is a disproportionate weight in favor of or against an idea or thing, usually in a way that is inaccurate, closed-minded, prejudicial, or unfair" (Steinbock, 1978). Biases can be innate or learned (Welsh and Begg, 2016). Propaganda can take many forms, including political speeches, advertisements, news reports, and social media posts (Guess and Lyons, 2020). Its goal is usually to influence people's attitudes and behaviors, either by promoting a particular ideology or by persuading them to take a specific action (Berinsky, 2017; Casavantes et al., 2024).

Hence our task [1], [2], addresses the critical need for analyzing bias and propaganda in multilingual political discourse. Developing annotation guidelines for complex data is a challenging task. In our tasks, we have identified the ideological bias related to support or against the government decisions on transgender rights, three language policies. We consider this as biases because it highly judgmental on sensitivity concerns. Annotating such bias is crucial for understanding how regional and linguistic identities are influenced in political discourse (Aksenov et al., 2021). The purpose of this annotation is to examine how political narratives shapes the public opinion by favoring or attacking specific policies and identities. To overcome these limitations, annotation efforts should incorporate diverse human perspectives: involving annotators from multiple cultural and linguistic backgrounds has been shown to reduce bias and capture nuanced interpretation. Addressing this challenge requires incorporating diverse perspectives and multi cultural insights during annotation, which can significantly enhance the robustness and fairness of NLP systems.

[1] https://sites.google.com/view/lt-edi-2025/tasks?authuser=0

[2] https://codalab.lisn.upsaclay.fr/competitions/22054

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 208–213

2 Related Works

Understanding and detecting bias in political discourse has become main concerns in computational social science (Heppell et al., 2023). Earlier research have explored the linguistic features of biased content and propaganda tactics in news articles, speeches and online comments (Lim et al., 2020; Allcott and Gentzkow, 2017). (Rashkin et al., 2017) analyzed linguistic pattern across different types of biased text, including fake news and political fact-checks. The work identified the subtle forms of bias through lexical and syntactical structures. (Da San Martino et al., 2019) offered a fine-grained taxonomy for identifying propaganda techniques in news articles. This work emphasized on detecting 18 specific propaganda techniques in news articles, such as appeal to fear, flag-waving, and loaded language. (Baly et al., 2020) presented study on predicting the political ideology of news articles. With a comprehensive dataset of with 34,737 news articles yielded the model's robustness. The authors suggested novel modeling approaches, such as a specially modified triplet loss function and adversarial media adaptation to deal with propaganda tactics in cultural contexts.

Multilingual bias detection presented by (Maity et al., 2024) created two large-scale datasets, mWikiBias and mWNC in eight Indian languages. The authors propose techniques for the detection of neutrality bias in politically and socially sensitive articles through models such as mDeBERTa and mT5. (Chavan and Kane, 2022) proposed a method for multi label propaganda detection using LLM. The WANLP 2022 shared task, which called for recognizing several propaganda strategies in a single text, inspired the development of their system. They achieved a micro-F1 score of 59.73% by using an ensemble of five models to handle the complexity of detecting 21 different propaganda techniques. (Zaghouani et al., 2024) conducted FIGNEWS shared task as a component of the ArabicNLP 2024 conference, which was held concurrently with ACL 2024. This work used the early stages of the Israel War on Gaza as a case study to examine bias and propaganda annotation in multilingual news posts. Their findings highlights the importance of clear guidelines and collaborative efforts in advancing NLP research on sensitive opinion analysis tasks.

3 Task Description

The shared task, addresses the crucial need for analyzing bias and propaganda in multilingual political discourse. This task aligns with the NLP community growing efforts to create datasets and guidelines for complex opinion analysis through collaborative shared tasks. There are two tasks in this shared task

- Task 1: Bias and Propaganda Annotation in English
  - Sub Task 1.1: US gender policy dataset The goal is to focuses on annotating contents related to Trump's US gender policy against transgender individuals. The task is to annotate based on the bias and propaganda guidelines in English texts that discuss this policy. There are totally 6 bias labels and 4 propaganda labels.
  - Sub task 1.2: Russia-Ukraine dataset Annotate the content of YouTube comments related to the Ukraine-Russia war in English. The task involves categorizing the comments based on bias and propaganda, following established guidelines for analyzing bias and propaganda in English texts. There are totally 8 bias labels and 4 propaganda labels.
- Task 2: Bias and Propaganda annotation in Tamil - Three language policy Dataset.
  The goal of task 2 is to provide annotating content related to the Three Language Policy/India's National Education Policy related issues. The task is to annotate based on the bias and propaganda guidelines in Tamil texts that discuss this policy. There are totally 7 bias labels and 4 propaganda labels.

Annotation Guidelines

- **Unbiased:** Neutral / Without favoritism.
  Example: "The US Gender Policy aims to address the rights of transgender individuals in military service, but the policy has been met with mixed reactions from different communities."
- **Biased Against US Gender Policy:** Criticizes negatively.

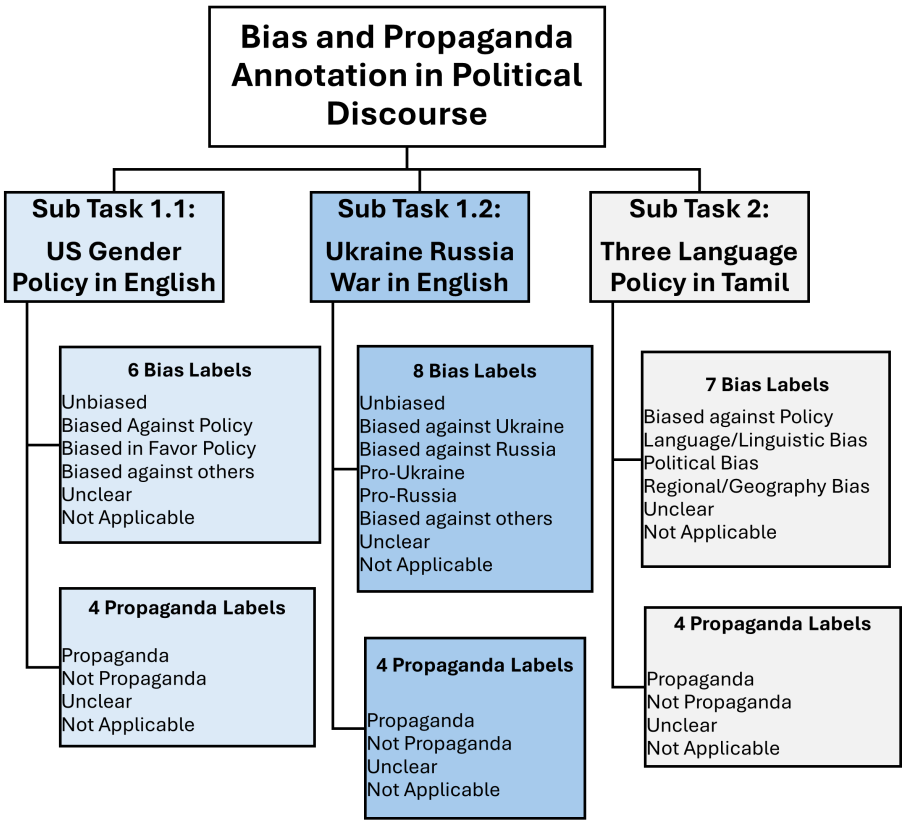

Figure 1: Task Overview and Annotation Labels

Example: "The US Gender Policy discriminates against transgender individuals by denying them the right to serve in the military, an unjust decision that harms the LGBTQ+ community."

- **Biased in Favor of US Gender Policy:** Strongly supports.
  Example: "Trump's US Gender Policy is a necessary measure to protect national security and uphold traditional values, and it's a step in the right direction for the country."
- **Biased Against Others:** Criticizes others negatively.
  Example: "Those who oppose Trump's US Gender Policy are out of touch with reality and are prioritizing political correctness over national security."
- **Unclear:** Text is incomplete.
  Example: "The debate over the US Gender Policy continues, but many people are still unsure about its long-term impact."
- **Not Applicable:** Irrelevant to the topic.
  Example: "The latest economic report shows growth in GDP this quarter."

4 Dataset Description

The dataset has been carefully curated from the YouTube platform by collecting comments from videos discussing the political concern. To identify relevant videos, we utilized a combination of hashtags such as *'National Education Policy'*, *'Three Language Policy'*, *'US gender policy'*, *'Ukraine Russia War'* alongside manual keyword searches including terms like *'Zelensky'*, *'Putin'* and *'Trump US Gender Policy'*.

After gathering the comments, we applied preprocessing steps to remove unrelated or noisy content. This included removing usernames, URLs, and comments containing fewer than three words to ensure data quality and relevance. The dataset statistics is shown in table 1

5 Participant Methodology

A total of 20 teams registered to participate in this shared task. However, only 2 teams submitted their

Task	No. of Samples	Vocab Size	Avg Length (in tokens)
Sub Task 1.1	7,911	9,475	21.54
Sub Task 1.2	5,099	12,619	37.95
Task 2	5,880	33,076	28.97

Table 1: Dataset statistics across different subtasks

results. Following are the detailed methodology of the participating teams.

- Scalar: Team Scalar contributed to Subtask 1.1 by performing manual annotations on the provided textual data, focusing on identifying bias and propaganda techniques present in discourse related to US gender policy. The annotation was carried out by two undergraduate students, both aged between 20–23. The manual annotation effort by Team Scalar is critical in generating high-quality, labeled datasets for training robust NLP models.

  Team Scalar also developed a transformer-based NLP model to detect both propaganda techniques and bias using the annotated data. They trained the model using transfer learning on a specially annotated dataset label encoding bias categories and tagging six propaganda techniques while adding extra non-propaganda examples to reduce class imbalance. The model was optimized with Adam and trained for four epochs (batch size 32) using sparse categorical crossentropy, achieving roughly 47.9 % accuracy.

- Mithun: This team present a context-aware neural model for detecting bias and propaganda in multilingual political discourse. Their approach stands out for its comprehensive annotation methodology, leveraging advanced metrics such as Bias Score, Cosine Similarity, Fairness Difference, and Weighted F1-score to evaluate both the fairness and accuracy of language models across diverse demographic and linguistic groups. By applying these metrics to English and Tamil datasets on sensitive topics like US gender policy, Ukraine/Russia discourse, and Three Language Policy the participant demonstrate significant disparities in model performance and fairness, highlighting the persistent challenges of bias in multilingual NLP.

6 Results and Discussion

Team Name (Sub Task 1.1)	Cohen's Kappa	Rank
Scalar	0.71	1
Mithun	0.39	2

Sub Task 1.2	Rank
Mithun (0.42)	1

Task 2: Tamil	Rank
Mithun (0.48)	1

Table 2: Bias and Propaganda Annotation Task results across English and Tamil subtasks.

Evaluation Metric: Cohen's Kappa

Cohen's Kappa (κ) is a statistical measure used to assess inter-annotator agreement for categorical classification tasks while correcting for chance agreement. It is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Where:

- P_o is the observed agreement between annotators,
- P_e is the expected agreement by random chance.

A κ value of:

- 1 indicates perfect agreement,
- 0 indicates chance-level agreement,
- Negative values indicate systematic disagreement.

Results and Quantitative Analysis

Table 2 presents the performance outcomes of participating systems across multiple subtasks:

- **Sub Task 1.1 (English)**: Team *Scalar* achieved the highest Kappa score of 0.71, indicating substantial agreement and strong classification capability. Team *Mithun* followed with 0.39, reflecting moderate agreement and highlighting difficulties in capturing nuanced propaganda techniques in English.
- **Sub Task 1.2 (English)**: Despite a moderate Kappa score of 0.42, *Mithun* ranked first, implying effective relative performance on this subtask.

- **Task 2 (Tamil)**: *Mithun* attained a Kappa of 0.48, leading the task. This score reflects moderate agreement in a low-resource language scenario, where annotation and detection challenges are more pronounced.

7 Conclusion

This shared task aims to enhance the annotation process of bias and propaganda in multilingual political discourse, focusing on English and Tamil texts. The shared task highlighted the role of clear guidelines, examples, and collaboration in advancing NLP research on complex, sensitive, and opinion analysis tasks. The resulting dataset and insights contribute valuable resources and direction for future work in this important area. Despite limited submissions, the task underscored the challenges in multilingual annotation and the importance of culturally-informed guidelines. Future efforts will focus on expanding the dataset, refining the annotation schema, and encouraging broader participation to build more generalizable models.

Acknowledgments

This work was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight).

References

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online. Association for Computational Linguistics.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Adam J Berinsky. 2017. Rumors and health care reform: Experiments in political misinformation. *British journal of political science*, 47(2):241–262.

Marco Casavantes, Manuel Montes-y Gómez, Luis Carlos González, and Alberto Barróñ-Cedeno. 2024. Propitter: A twitter corpus for computational propaganda detection. In *Advances in Soft Computing*, pages 16–27, Cham. Springer Nature Switzerland.

Tanmay Chavan and Aditya Manish Kane. 2022. ChavanKane at WANLP 2022 shared task: Large language models for multi-label propaganda detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 515–519, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Pratim Datta, Mark Whitmore, and Joseph K. Nwankpa. 2021. A perfect storm: Social media news, psychological biases, and ai. *Digital Threats*, 2(2).

Andrew M. Guess and Benjamin A. Lyons. 2020. *Misinformation, Disinformation, and Online Propaganda*, page 10–33. SSRC Anxieties of Democracy. Cambridge University Press.

Freddy Heppell, Kalina Bontcheva, and Carolina Scarton. 2023. Analysing state-backed propaganda websites: a new dataset and linguistic study. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5729–5741, Singapore. Association for Computational Linguistics.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Ankita Maity, Anubhav Sharma, Rudra Dhar, Tushar Abhishek, Manish Gupta, and Vasudeva Varma. 2024. Multilingual bias detection and mitigation for Indian languages. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 24–29, Torino, Italia. ELRA and ICCL.

Preslav Nakov and Giovanni Da San Martino. 2020. Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–19, Online. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017*

Conference on Empirical Methods in Natural Language Processing, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Bonnie Steinbock. 1978. Speciesism and the idea of equality. *Philosophy*, 53(204):247–256.

Matthew Welsh and S. Begg. 2016. What have we learnt? insights from a decade of bias research. *Australian Petroleum Production and Exploration Association Journal*, 56.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa El-Beltagy, and Muhammed AbuOdeh. 2024. The FIGNEWS shared task on news media narratives. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 530–547, Bangkok, Thailand. Association for Computational Linguistics.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):e0150989.

Findings of the Shared Task on Caste and Migration Hate Speech Detection

Saranya Rajiakodi[1], Bharathi Raja Chakravarthi[2], Rahul Ponnusamy[2], Shunmuga Priya MC[2], Prasanna Kumar Kumaresan[2], Sathiyaraj Thangasamy[3], Bhuvaneswari Sivagnanam[1], Balasubramanian Palani[4], Kogilavani Shanmugavadivel[5], Abirami Murugappan[6], Charmathi Rajkumar[7]

[1] Central University of Tamil Nadu, India

[2] School of Computer Science, University of Galway, Ireland

[3]Department of Tamil, Sri Krishna Adithya College of Arts and Science, Tamil Nadu, India

[4]Indian Institute of Information Technology Kottayam, Kerala, India [5]Kongu Engineering College, Tamil Nadu, India

[6]Department of Information Science and Technology, Anna University, India [7]The American College, Tamil Nadu, India

Abstract

Hate speech targeting caste and migration communities is a growing concern in online platforms, particularly in linguistically diverse regions. By focusing on Tamil language text content, this task provides a unique opportunity to tackle caste or migration related hate speech detection in a low resource language Tamil, contributing to a safer digital space. We present the results and main findings of the shared task caste and migration hate speech detection. The task is a binary classification determining whether a text is caste/migration related hate speech or not. The task attracted 17 participating teams, experimenting with a wide range of methodologies from traditional machine learning to advanced multilingual transformers. The top performing system achieved a macro F1-score of 0.88105, enhancing an ensemble of fine-tuned transformer models including XLM-R and MuRIL. Our analysis highlights the effectiveness of multilingual transformers in low resource, ensemble learning, and culturally informed socio political context based techniques.

1 Introduction

In the current digital world, the online social network platforms such as Twitter, YouTube, LinkedIn, Facebook and WhatsApp are widely used by the individuals from the different parts of the country (Kruse et al., 2018; Kubin and von Sikorski and, 2021). Due to the high-speed internet facility, the news like an offensive speech (Sreelakshmi et al., 2024), fake news (Subramanian et al., 2025), hate speech and its curated forms against the caste or people are disseminated across the globe in a fraction of minutes (Rajiakodi et al., 2024; Chakravarthi et al., 2025). Hence, it leads to lot of issues such as protests, inflation in the economy of the country and majorly affects the healthy environment in the society (Jost et al., 2018). The hate speech and offensive words are more important to eradicate from the society to preserve the equality among the citizens. Apart from the global issues, the discrimination and bias against the certain communities and its people will affect directly their communities and their people in terms of their mental health (Abubakar et al., 2022; Matamoros-Fernández and Farkas, 2021). Hence, the hate speech against any caste/migration is an important issue and it should be detected to save the society and nation's peace.

India is a democratic country and follows the "Unity in Diversity". According to the Constitution of India [1], Articles 14-18 deal with the right to equality. Hence if any people discriminate others in terms of language, caste or any variants, the government will take immediate actions against them with these articles. With the proliferation of social media platforms, public discourse has become increasingly dynamic, often marked by a surge of user-generated content in response to specific events (Shanmugavelan, 2022). This environment has made the manual identification of caste-based hate speech both time-consuming and resource-intensive. To address this challenge, there has been a growing adoption of Natural Language Processing (NLP) based systems capable of automatically detecting hate speech across multiple languages and platforms, thereby enhancing the efficiency and scalability of monitoring efforts (Roy et al., 2022; Sharma et al., 2025).

Researchers are developed various machine learning (ML) and deep learning (DL) model to identify offensive language and hate speech from the various regional low-resource languages such as Hindi (Rani et al., 2020; Rajak and Baruah, 2025), Telugu, Malayalam and Tamil (G et al., 2025). A novel dataset which comprises of hate speech by castes and migration collected from

[1]https://bit.ly/44tqpwg

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 214–220

YouTube and published in LT-EDI@EACL 2024 (Rajiakodi et al., 2024). In this shared task, the team of participants have developed an automatic hate speech detection against the caste and migration model.

Specifically, the top ranking teams used an ensemble model which consists of transformer based multilingual models such as XLM-RoEERTa and MuRIL for contextual embedding of sentences to predict hate speech that achieves higher macro F1 score among other models. The work by (Alam et al., 2024) developed six different ML models and three DL models including Bi-LSTM, Attention, Bi-LSTM-CNN to identify the hate or not-hate speech. Finally, they designed transformer-based models such as M-BERT, XLM-R and Tamil-BERT for effective contextual word embedding, which achieves better performance than the other models.

In the subsequent sections, we have discussed the task description, dataset statistics, methodologies used by the participants to detect caste/migration hate speech in Tamil, and their results and ranking.

2 Tasks and Dataset

2.1 Task Description

The task's goal is to classify YouTube text comments in Tamil into two categories: hate speech on caste/migration, non-hate speech on caste/migration. Participants were provided with:

- **Training and Development Sets**: These sets were annotated with labels to allow participants to train and fine tune their models effectively.
- **Testing Set**: This set was unlabeled, requiring participants to generate predictions without the aid of ground truth labels, which were reserved for evaluation purposes.

2.2 Dataset Description

The dataset has been carefully curated from the YouTube platform by collecting comments from videos discussing caste and migration related issues. To identify relevant videos, we utilized a combination of hashtags such as *'vadakan'* (North Indian), *'devar'* (a caste), alongside manual keyword searches including terms like *'Agamudiyar'* (a caste group) and *'Melpathi temple issue'* (a specific caste related controversy).

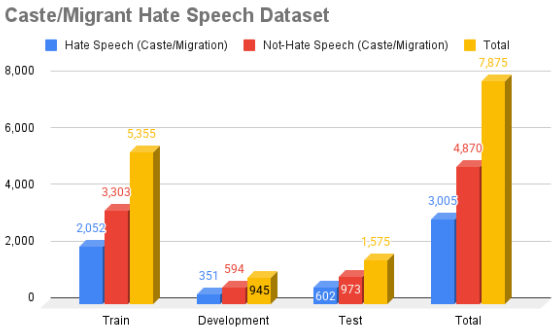

Figure 1: Dataset Statistics

After gathering the comments, we applied preprocessing steps to remove unrelated or noisy content. This included removing usernames, URLs, and comments containing fewer than three words to ensure data quality and relevance.

Each comment was then annotated independently by three annotators, all proficient in the Tamil language. The annotation team was deliberately heterogeneous, composed of individuals from diverse caste backgrounds—including historically marginalized and dominant groups various age ranges from young adults (20–29 years) to middle aged adults (30–45 years), as well as gender and professional diversity. This diversity in the annotator pool was intended to reduce bias and enhance the reliability of the annotations.

The annotation labels are defined as follows:

- **Hate speech on caste/migration:** Comments containing abusive, disrespectful, or discriminatory language, including ridicule or mockery, and content aimed at delegitimizing specific caste or migrant groups.
- **Not-Hate speech on caste/migration:** Comments that do not contain any discriminatory or derogatory remarks towards individuals based on their caste or migration status.

The reliability of the annotations was measured using Krippendorff's alpha, which yielded a high agreement score of 0.83441, indicating strong consistency among annotators. The detailed statistics of the dataset is shown in Figure 1.

3 Results

The evaluation metric used was the macro-averaged F1-score, calculated using the scikit-learn library[2].

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.

3.1 Participant Statistics

Our competition was created on Codalab [3], and has attracted registrations and a total of 17 submissions. Team Samsung Research CUET_N317 won first place in the task, holding a significant lead over the second place team. Teams CUET's_white_walkers and Wise respectively captured the second and third places. The official leaderboards for this task is shown in Table 1.

3.2 Participant Methodology

In this section, we first summarize common features for all teams based on the information they provided in the Google Documents. Then, we delve into the methods employed by the top 5 teams, accompanied by brief descriptions of the approaches utilized. The approaches of all other teams are presented briefly in table 1.

- **CUET_N317 (Md. Nur Siddik Ruman, 2025)**: This team employed three distinct strategies: traditional machine learning models with TF-IDF features, individual fine-tuned multilingual transformers (e.g., XLM-R, MuRIL, IndicBERT), and an ensemble of five fine-tuned multilingual transformer models. The ensemble model achieved the best performance. Extensive hyperparameter tuning and early stopping techniques were used to optimize results.

- **CUET's_White_Walkers (Jidan Al Abrar, 2025)**: The team used a fine-tuned TamilBERT model, incorporating techniques such as cosine annealing and the AdamW optimizer. They emphasized training stability and leveraged visualizations to analyze performance. The model was optimized for Tamil text without relying on transliteration or external datasets.

- **Wise (Ganesh Sundhar S, 2025)**: They explored both transliterated and original Tamil text preprocessing. A hybrid approach combined TF-IDF, Truncated SVD, MLP, and multilingual transformer embeddings. Feature fusion and late decision fusion were used to aggregate predictions. The team emphasized the comparative value of transliteration-aware vs. native Tamil models.

- **CUET_blitz_aces**: This team adopted a three-phase pipeline: initial training with traditional ML (TF-IDF + classifiers), fine-tuning multilingual transformers, and a final ensemble voting mechanism. They conducted thorough evaluation and analysis at each stage to select optimal model configurations.

- **hinterwelt (MD AL AMIN, 2025)**: The team experimented with XLM-R Large, MuRIL-large-cased, and IndicBERT models. Training included advanced strategies like learning rate schedulers, gradient clipping, and early stopping. The final system combined multiple transformer models, optimized with rigorous experimentation for multilingual hate speech detection.

- **ItsAllGoodMan (Amritha Nandini KL, 2025)**: Combined train/dev data, used back translation for data augmentation, segmented hashtags, replaced mentions, and converted emojis. Tried TF-IDF with Random Forest, soft voting ensembles with various embeddings and ML models, and ensembles of fine-tuned transformers.

- **NS**: Transliterated Tamil text to English and vice versa for consistent embeddings. Trained six ML models on different embedding sets and selected top three (XGBoost, Logistic Regression and MLP) for final predictions.

- **CUET_perceptrons** Trained several transformer-based models and combined their predictions using a weighted ensemble, where better-performing models had higher influence on the final label.

- **KCRL**: Combined TamilBERT and distilbert-base-multilingual-cased with k-fold cross-validation. Used a classification architecture that concatenated CLS token, mean pooling, and max pooling.

- **Cuet_try_NLP**: Used a traditional ML pipeline with Bag of Words (unigrams, bigrams) and Random Forest for classification. Focused on computational efficiency but lacked semantic understanding.

- **Wictory**: Hybrid approach: fine-tuned MuRIL-based transformer with discriminative learning rates and dense layers, plus multilingual embeddings fed into an SVM. Combined

html

[3] https://codalab.lisn.upsaclay.fr/competitions/21884

Rank	Team Name	Run	Macro F1 Score	Highlighted Methodology
1	**CUET_N317** (Md. Nur Siddik Ruman, 2025)	2	**0.88105**	Fine-tuned multilingual transformer model
2	CUET's_white_walkers (Jidan Al Abrar, 2025)	1	0.86289	Fine tuned **Tamil-BERT** with cosine annealing, AdamW, and performance visualization.
3	Wise (Ganesh Sundhar S, 2025)	1	0.81827	Transliteration vs Non-transliteration preprocessing, TF-IDF + Truncated SVD + MLP + Transformer fusion.
4	CUET_blitz_aces (Shahriar Farhan Karim, 2025)	3	0.81682	3-phase: ML (TF-IDF), transformer fine-tuning, and ensemble voting of 5 models.
5	hinterwelt (MD AL AMIN, 2025)	2	0.80916	Fine-tuned **XLM-R Large**, MuRIL-large-cased, IndicBERT with LR scheduler, gradient clipping, early stopping.
6	ItsAllGoodMan (Amritha Nandini KL, 2025)	3	0.80364	TF-IDF + Random Forest (best), back translation for augmentation, MuRIL + voting ensemble.
7	NS (Nishanth S, 2025)	1	0.80095	Tamil English transliteration embeddings + XGBoost, Logistic Regression, MLP.
8	CUET_perceptrons	2	0.79812	Weighted ensemble of multiple transformer models.
9	KCRL	1	0.79081	Triple embedding (CLS, mean, max) + TamilBERT and DistilBERT with cross-validation.
10	Cuet_try_NLP	2	0.78175	Bag-of-Words + n-grams + Random Forest.
11	Wictory	1	0.76630	MuRIL + SVM hybrid with focal loss, dense layers, class weighting.
12	Solvers (Mohanapriya K T, 2025)	2	0.76518	CNN-based model + BERT fine-tuning.
13	DravLang	1	0.76182	BiLSTM + Attention + Voting and Stacking Classifiers (XGBoost, SVM, NB).
14	girlsTeam (Towshin Hossain Tushi, 2025)	1	0.74522	MuRIL/mBERT/IndicBERT + BiLSTM + ML classifiers + multilingual embeddings.
15	ScalarLab	1	0.73308	M-BERT and XLM-R; ensemble predictions from both.
16	SSN_IT (Maria Nancy C, 2025)	1	0.72462	BERT multilingual model + preprocessing Tamil and Romanized Tamil.
17	EM-26 (Tewodros Achamaleh, 2025)	1	0.65672	XLM-Roberta + AdamW

Table 1: Leader board Results with Methodologies

outputs for interpretability and deep contextual understanding.

- **Solvers (Mohanapriya K T, 2025)**: Explored deep learning models starting with CNNs, then modified for multi-class with softmax. Also fine-tuned a BERT-based model for contextual understanding and compared their effectiveness.
- **DravLang**: Hybrid approach: preprocessed text with tokenization and TF-IDF, trained a BiLSTM with Attention, an ensemble Voting Classifier (XGBoost, SVM, Naïve Bayes), and a Stacking Classifier (XGBoost, SVM, Logistic Regression meta-learner)
- **girlsTeam (Towshin Hossain Tushi, 2025)**: Comprehensive hybrid: transformer-based embeddings (MuRIL, mBERT, IndicBERT), enhanced with BiLSTM, and traditional ML classifiers (Random Forest, SVM, XGBoost) using TF-IDF, FastText, and Word2Vec. Included tokenization, normalization, transliteration, and class balancing.
- **ScalaraLab**: Used Multilingual BERT and XLM-R models, then combined their outputs for predictions.
- **SSN_IT (Maria Nancy C, 2025)**: Fine-tuned bert-base-multilingual-cased on Tamil and Romanized Tamil. Preprocessing included lowercasing, URL/special character removal, and BERT tokenization. Used AdamW optimizer, cross-entropy loss, and evaluated with standard metrics.
- **EM-26 (Tewodros Achamaleh, 2025)**: Fine-tuned XLM-RoBERTa , with class weighting and augmentation (word swapping/dropping). Used AdamW, learning rate scheduler, and early stopping based on F1 score.

The leaderboard results and methodological descriptions reveal a clear trend toward the effectiveness of multilingual transformer-based models for caste and migration-related hate speech detection. Teams that employed fine-tuned transformer architectures such as XLM-R, MuRIL, and IndicBERT often in ensemble settings consistently outperformed traditional machine learning approaches. The top-performing submissions integrated advanced training strategies, including learning rate schedulers, gradient clipping, early stopping, and ensemble voting, to boost performance and ensure robustness. Additionally, teams that conducted rigorous ablation studies, language-specific preprocessing, and hybrid model experimentation (eg: combining TF-IDF with deep models) demonstrated a deeper understanding of the

multilingual and culturally nuanced nature of the task. Overall, the results highlight that a blend of domain-specific linguistic preprocessing and transformer-based modeling yields the most competitive outcomes in multilingual hate speech detection tasks.

4 Conclusion

This shared task on caste and migration hate speech detection focused on detecting hate speech against people based on their caste or migration status, especially in the Tamil language. Since this type of hate is often ignored in existing AI systems, this task helped researchers and developers create models that are more aware of these social issues. Participants used many different methods, including machine learning and advanced models like BERT and XLM-R, to analyze the text. The best teams used a mix of good text cleaning, careful model training, and techniques like ensemble voting and data balancing to get better results. The top system scored a high macro F1 of 0.88105, showing that with the right tools and approaches, it is possible to detect even subtle and hidden forms of hate speech. This task showed that we need models that understand regional languages and cultural context to handle online hate more effectively. It also highlights the need for more work in areas like explaining model predictions, creating respectful counter-speech, and labeling more detailed types of hate.

Acknowledgments

Author Bharathi Raja Chakravarthi was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight). This work is also supported by the Centre for Research Training in Artificial Intelligence grant number SFI/18/CRT/6223, as well as a grant from the College of Science and Engineering, University of Galway, Ireland.

References

Ibrahim Abubakar, Lu Gram, Sarah Lasoye, E. Tendayi Achiume, Laia Becares, Gurpreet Kaur Bola, Rageshri Dhairyawan, Gideon Lasco, Martin McKee, Yin Paradies, Nidhi S. Sabharwal, Sujitha Selvarajah, Geordan Shannon, and Delan Devakumar. 2022. Confronting the consequences of racism, xenophobia, and discrimination on health and health-care systems. *The Lancet*, 400(10368):2137–2146. PMID: 36502851.

Md Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_Manning@LT-EDI 2024: Transformer-based approach on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 238–243, St. Julian's, Malta. Association for Computational Linguistics.

Giri Prasath R Anerud Thiyagarajan Sachin Kumar S Amritha Nandini KL, Vishal S. 2025. Itsallgoodman@lt-edi-2025: Fusing tf-idf and muril embeddings for detecting caste and migration hate speech. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

B. R. Chakravarthi, S. Rajiakodi, R. Ponnusamy, Bhuvaneswari Sivagnanam, Sara Yogesh Thakare, and Sathiyaraj Thangasamy. 2025. Detecting caste and migration hate speech in low-resource tamil language. *Language Resources and Evaluation*.

Jyothish Lal G, Premjith B, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Ratnavel Rajalakshmi. 2025. Overview of the shared task on multimodal hate speech detection in Dravidian languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 114–122, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Gnanasabesan G Hari Krishnan N Dhanush MC Ganesh Sundhar S, Durai Singh K. 2025. Wise@ltedi2025: Combining classical and neural representations with multiscale ensemble learning for codemixed hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Ariful Islam Md Mehedi Hasan Md Mubasshir Naib Mohammad Shamsul Arefin Jidan Al Abrar, Md Mizanur Rahman. 2025. Cuets_white_walkers@lt-edi 2025: Transformerbased model for the detection of caste and migration hate speech. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

John T. Jost, Pablo Barbera, Richard Bonneau, Max Langer, Miriam Metzger, Jonathan Nagler, Josh Sterlin, and Joshua A. Tucker. 2018. How social media facilitates political protest: Information, motivation, and social networks. *Advances in Political Psychology*, 39(S1):85–118.

Lisa M. Kruse, Dawn R. Norris, and Jonathan R. Flinchum and. 2018. Social media as a public sphere? politics on social media. *The Sociological Quarterly*, 59(1):62–84.

Emily Kubin and Christian von Sikorski and. 2021. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206.

Swathika R Maria Nancy C, Radha N. 2025. Ssn_it_hate@ltedi2025: Caste and migration hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

Md. Abdur Rahman Md Sajid Hossain Khan Md Ashiqur Rahman MD AL AMIN, Sabik Aftahee. 2025. Hinterwelt@lt-edi 2025: A transformer-based detection of caste and migration hate speech in tamil social media. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Hasan Murad Md. Nur Siddik Ruman, Md. Tahfim Juwel Chowdhury. 2025. Cuet_n317@lt-edi2025: Detecting hate speech related to caste and migration with transformer models. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Devasri A Bharath P Ananthakumar S Mohanapriya K T, Anirudh Sriram K S. 2025. Solvers@lt-edi-2025: Caste and migration hate speech detection in tamil-english code-mixed text. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Sachin Kumar S Nishanth S, Shruthi Rengarajan. 2025. Ns@ltedi2025 caste migration based hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Santosh Rajak and Ujwala Baruah. 2025. A comprehensive hindi hostile post detection dataset: Aannotated resource for fine-grained hostility analysis on twitter posts in the hindi language. *Expert Systems with Applications*, 289:128191.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 145–151, St. Julian's, Malta. Association for Computational Linguistics.

Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 42–48.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

Hasan Murad Shahriar Farhan Karim, Anower Sha Shajalal Kashmary. 2025. Cuet_blitz_aces@lt-edi-2025: Leveraging transformer ensembles and majority voting for hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Murali Shanmugavelan. 2022. Caste-hate speech and digital media politics. *Journal of Digital Media amp; Policy*, 13(Special Issue: '(Re)Iterations, Transgressions, Recognition: Politics and Practices of Media Policies in South Asia'):41–55.

Deepawali Sharma, Vedika Gupta, Vivek Kumar Singh, and Bharathi Raja Chakravarthi. 2025. Stop the hate, spread the hope: An ensemble model for hope speech detection in english and dravidian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach. *IEEE Access*, 12:20064–20090.

Malliga Subramanian, Premjith B, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the shared task on fake news detection in Dravidian languages-DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 759–767, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Mikiyas Mebiratu Sara Getachew Grigori Sidorov Tewodros Achamaleh, Abiola T. O. 2025. Em-26@lt-edi 2025: Caste and migration hate speech detection in tamil-english code-mixed social media texts. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Rehenuma Ilman Samia Rahma Towshin Hossain Tushi, Walisa Alam. 2025. girlsteam@lt-edi-2025:

Caste/migration based hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Overview of the Shared Task on Detecting Racial Hoaxes in Code-Mixed Hindi-English Social Media Data

Bharathi Raja Chakravarthi[1], Prasanna Kumar Kumaresan[1], Shanu Dhawale[1], Saranya Rajiakodi[2], Sajeetha Thavareesan[3], Subalalitha Chinnaudayar Navaneethakrishnan[4], Durairaj Thenmozhi[5]

[1]School of Computer Science, University of Galway, Ireland
[2]Central University of Tamil Nadu, India
[3]Department of Computing, Eastern University, Sri Lanka
[4]Department of Computer Science & Engineering,
SRM Institute of Science and Technology, Tamil Nadu, India
[5]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
Correspondence: bharathi.raja@universityofgalway.ie

Abstract

The widespread use of social media has made it easier for false information to proliferate, particularly racially motivated hoaxes that can encourage violence and hatred. Such content is frequently shared in code-mixed languages in multilingual nations like India, which presents special difficulties for automated detection systems because of the casual language, erratic grammar, and rich cultural background. The shared task on detecting racial hoaxes in code-mixed social media data aims to identify the racial hoaxes in Hindi-English data. It is a binary classification task with more than 5,000 labeled instances. A total of 11 teams participated in the task, and the results are evaluated using the macro-F1 score. The team that employed XLM-RoBERTa secured the first position in the task.

1 Introduction

In today's world, social networks play a vital role in how people get their information, but they also make it easier for false claims to spread quickly (Chakravarthi, 2024; Lopez, 2022). One serious kind of fake story is racial hoaxes, where people wrongly blame a particular person or community for a crime or incident simply because of their race, religion, or background. What makes these hoaxes really dangerous is that they create and spread unfair stereotypes, divide people, and can create real trouble, such as fights or community clashes. The problem becomes even harder to deal with in a diverse country like India, where people often mix languages like Hindi and English in their on-line posts, so even tech tools have a hard time spotting these misleading claims.

To take a step toward solving this problem, a shared task called "Detecting Racial Hoaxes in Code-Mixed Hindi-English Social Media Data" was launched as part of the LT-EDI Workshop 2025. As part of this task, a new dataset called HoaxMixPlus was introduced. It includes 5,105 YouTube comments written in a mix of Hindi and English, each carefully marked to show whether it contains a racial hoax. This data set reflects how complex and sensitive language can be when people of different backgrounds talk online. Unlike general fake news detection(Sivanaiah et al., 2022; Katariya et al., 2022; De et al., 2021), this task focuses on more subtle signs such as blaming someone without proof, hinting at accusations without saying them outright, or using words that carry hidden meaning. Spotting these patterns requires systems that can detect deeper language cues and the context behind the words.

The primary goal of the task is to benchmark and encourage the development of robust classification models capable of identifying racial hoaxes in a low-resource and highly informal setting. The availability of HoaxMixPlus serves not only as a starting point for building such models but also as a valuable contribution to the field of computational social science, particularly in understanding and mitigating the spread of race-based misinformation in South Asian digital ecosystems. Participants were expected to take a step toward tackling this issue related to code switching, informality, implicit bias, and contextual ambiguity.

11 teams from universities and research labs participated in the shared task, trying a variety of methods - from simple machine learning models to more advanced systems such as transformers, multilingual word embeddings, and fine-tuned large language models. This overview paper brings together a summary of the dataset, how the task was de-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 221–227

signed, how the results were measured, and the different techniques used by the teams. The insights gained from this task can help move the field forward, especially in creating more ethical and culturally sensitive language tools to spot misinformation in code-mixed, low-resource settings.

2 Related Work

In recent years, the rise of social networks has led to an explosion in code-mixed communication, especially in regions that are linguistically diverse like India. Hindi, an official and widely spoken language in North India (Srivastava et al., 2020), is traditionally written in Devanagari script. However, informal online interactions have adopted a code-mixed setup in which users express their personal thoughts and emotions. These pose significant challenges for automated detection of abusive language and hate speech. Traditional corpora and models, trained on monolingual and structured data, often fall short when faced with the informal grammar, intentional misspellings, and hybrid syntax typical of code-mixed texts (Vijay et al., 2018; Kumar et al., 2018; Nayak and Joshi, 2022; Dey and Fung, 2014). Detecting nuanced social phenomena like hate speech, especially when it is unstructured code-mixed language, more than just large dataset it requires deeply annotated corpora, careful curation, and time-intensive fact-checking processes.

Although some progress has been made in this direction—for example, Italian social media texts (Bosco et al., 2023), multilingual datasets in Italian, Spanish, and French annotated in racial hoax for immigrant stereotype (Bourgeade et al., 2023), English dataset named StereoSet (Nadeem et al., 2021) are some of the novel corpus. However, access to similarly robust and diverse datasets for Indian code-switched scenarios remains limited. As a result, researchers working in this area face considerable challenges not only in developing accurate detection models, but also in constructing the foundational datasets needed to reflect the complex socio-linguistic realities of code-mixed data and bias online.

The development of Hindi-English code-mixed datasets for abusive language and hate speech detection presents significant challenges, as outlined in prior studies (Bohra et al., 2018; Kumar et al., 2018; Nayak and Joshi, 2022; Dey and Fung, 2014). Unlike general hate speech, racial hoaxes demand fine-grained analysis of implicit and explicit biases targeting specific social groups. Transformer-based models have been applied in related contexts, such as Bangla-English sentiment analysis on YouTube (Kar and Jana, 2024), and Swahili-English political misinformation (Amol et al., 2024). Approaches such as BiLSTM-CRF architectures (Bhattu et al., 2020) and transformer-based models like HingBERT, HingRoBERTa and HingGPT—pre-trained (Nayak and Joshi, 2022) have shown promise for POS tagging and sentiment classification in code-mixed data. Novel language augmentation strategies, including word-level interleaving and Latin-script resource integration (Sharma et al., 2022; Takawane et al., 2023), further enhance classification performance combined with dedicated lexicon for Latin-script Hindi-English words, combined with fine-tuned Multilingual BERT and MuRIL models, effectively tackles the unique challenges presented by Hindi-English code-mixed datasets. These innovations collectively highlight the evolving methodologies and underscore the importance of tailored pre-training and annotation in handling the complexity of Hindi-English code-mixed text.

3 Task Description

The Racial Hoax Detection task, featured at LT-EDI@LDK 2025[1], challenges participants to develop automated systems capable of identifying racial hoaxes within code-mixed Hindi-English social media content. Racial hoaxes refer to false statements that use misleading information to falsely accuse people or groups because of their social, ethnic, or religious backgrounds, including caste, nationality, or religion. The task involves classifying user-generated comments, primarily sourced from YouTube, into two categories: those containing racial hoaxes (positive) and non-racial hoaxes (negative). The binary classification process becomes challenging because the data contains code-switching, informal language, and sociopolitical complexities. Participants are provided with a labeled dataset and are expected to submit predictions on a held-out test set. The evaluation metric for this task is the macro-averaged F1 score, which balances performance across both classes. The research goal of this task focuses on developing multilingual and low-resource natural language processing while prioritizing ethical AI applications for misinformation detection.

[1] https://codalab.lisn.upsaclay.fr/competitions/21885

4 Dataset Descriptions

The HoaxMixPlus dataset consists of 5,105 YouTube comments that are code-mixed Hindi-English to help detect and classify racial hoaxes in social media discourse. Racial hoaxes refer to false or fabricated claims that target people or groups based on their social or ethnic identity, such as caste, religion, or nationality. The dataset was created by scraping over 210,000 YouTube comments from socio-politically relevant videos using targeted keywords such as Dalit, CAA-NRC, Manipur, Rohingyas, Khalistan, and Kerala Story. We selected these keywords to ensure a diverse collection of comments that accurately reflected real-world discourse on potentially sensitive topics. The Polyglot language detection library performed two functions: it eliminated code-mixed Hindi-English comments, and it eliminated all non-Hindi-English code-mixed data and monolingual content. The comments received binary classification labels of Positive (Racial Hoax) for false accusations and fabricated stories, and stereotypes against social group,s and Negative (Non-Racial Hoax) for neutral, factual, or non-misleading content.

Table 1: Data Distribution of HoaxMixPlus

Class	Train	Validation	Test
Racial Hoax (Positive) state	741	247	247
Non Racial Hoax (Negative) state	2,320	775	775
Total	3,061	1,022	1,022
Total Dataset		5,105	

Native Hindi-speaking annotators performed the annotations while using a custom GPT-4 chatbot for consistency and guidance. The team resolved ambiguous cases by reaching majority consensus, and Krippendorff's alpha ($\alpha = 0.747$) measured the inter-annotator agreement at substantial levels. The dataset contains different code-switching patterns, which include intra-sentential switching (within a sentence), inter-sentential switching (between sentences), and tag switching (inserting single words or short phrases from another language), with most comments written in Latin script using Hindi grammar and English vocabulary or vice versa. Table 1 shows that the final dataset contains 5,105 comments with 152,250 tokens and 17,314 unique tokens after removing emojis and URLs and filtering out short and overly long comments. The dataset contains 3,061 training examples and 1,022 validation and 1,022 test examples, which maintain the label distribution. The dataset received performance enhancement through transliteration to Devanagari script and language tagging (EN, HI, OOV) and lexicon-based spelling correction and disambiguation using a curated racial hoax knowledge base.

5 Methodology

The Racial Hoax Detection task participants employed different methods to solve the problem while working toward precisely identifying racially motivated misinformation in noisy Hindi-English social media comments. The task required systems to process code-switching linguistic complexity together with the delicate sociopolitical elements that frequently appear in racial hoaxes. The systems needed to maintain strong natural language understanding capabilities while being sensitive to cultural details, which standard sentiment and hate speech detection pipelines typically miss. The binary classification framework of the task required models to achieve both high precision and reliability because false positives and false negatives had significant real-world consequences.

The majority of teams began their architecture with advanced multilingual transformer models. The team selected XLM-RoBERTa, MuRIL, mBERT, and DeBERTa models because these models showed excellent cross-lingual performance and worked well in low-resource conditions. The preprocessing pipelines adapted the input data through techniques that included Devanagari script transliteration and EN, HI, and OOV language tagging and stopword handling, and social media noise removal of emojis, URLs, and usernames. The systems either concentrated on syntactic cleaning or used semantic features and handcrafted lexicons to detect hoax-related patterns. The team applied data augmentation techniques, which included synonym replacement, back-translation, and adversarial examples, to enhance model generalizability and reduce the imbalance between hoax and non-hoax examples. The distinguishing factor between submissions was their effective combination of linguistic heuristics with ensemble strategies and domain-specific knowledge bases.

The **CUET's_White_Walkers** (Rahman et al., 2025b) team established their system using XLM-RoBERTa as one of its key contributions. The team started by freezing the lower transformer layers during the initial training period to maintain general language representations before fine-tuning the

Table 2: Ranklist of Hindi-English

S.No.	Team Name	Run	macro F1	Rank
1	CUET's_White_Walkers (Rahman et al., 2025b)	2	0.75	1
2	Hope_for_best (Yadav et al., 2025)	-	0.72	2
3	KCRL	-	0.71	3
4	HoaxTerminators (Rabbani et al., 2025)	3	0.70	4
5	Hinterwelt (Rahman et al., 2025a)	1	0.69	5
6	Belo Abhigyan	-	0.68	6
7	KEC-Elite-Analysts (Subramanian et al., 2025)	1	0.68	6
8	Dll5143A (Yadav and Singh, 2025)	3	0.67	7
9	EM-26 (Achamaleh et al., 2025)	-	0.63	8
10	Squad	1	0.58	9
11	CVF_NITT	-	0.43	10

model on HoaxMixPlus. The team implemented learning rate scheduling and early stopping techniques to prevent the model from overfitting on the unbalanced dataset. The **Hope_for_best** (Yadav et al., 2025) team concentrated their strategy on MuRIL which represents a transformer model trained on Indian languages. The preprocessing began with an intense method to clean and normalize the code-mixed text. The model received transliteration and language tagging to standardize the inputs before undergoing stratified class balancing training. The model concentrated on preserving stability against social media noise and unclear sentence structures.

The team **EM-26** (Achamaleh et al., 2025) employed mBERT's multilingual features in their approach. The team employed traditional fine-tuning together with lexicon-aware normalization in their approach. The team focused on Hindi words with inconsistent spellings and transliterations through normalization techniques and token filtering to prepare training inputs. The team studied multiple approaches to control model complexity through dropout and attention masking techniques. **CVF_NITT** implemented an ensemble-based architecture through weighted softmax aggregation of XLM-R, mBERT, and MuRIL outputs. The ensemble received advantages from a strong preprocessing pipeline, which eliminated emojis and expanded contractions and standardized slang. They also implemented a language-model-agnostic token weighting scheme to amplify social cues linked with misinformation.

KCRL introduced a multi-model ensemble that merged predictions from XLM-RoBERTa, MuRIL, and DeBERTa. Their work emphasized cross-verification through handcrafted stereotype dictionaries and explored variations in token-level input formatting. They also curated a preprocessing module to identify and standardize acronyms, which are often misleading or misused in racial hoax discourse. The team **Hoax Terminators** (Rabbani et al., 2025) submitted three variations of their system. Their first run combined MuRIL and XLM-R in a straightforward ensemble using language-tagged inputs and transliterated text. The second run used a DeBERTa-based model, which incorporated a racial hoax lexicon to introduce inductive bias during learning. The third run used mBERT with strong data augmentation techniques such as synonym injection and entity replacement to simulate social media noise and test robustness. The team **Dll5143A** (Yadav and Singh, 2025) examined three runs through the BaCoHoax framework. The first configuration used a MuRIL-based backbone trained on clean, length-filtered samples from HoaxMixPlus. The second version incorporated interleaved language tags at the word level and leveraged linguistic cues via attention masking. Their final submission utilized DeBERTa, coupled with character-level embeddings and a stereotype term dictionary that informed contextual weightings during classification.

Altogether, these varied methodologies illustrate how teams tailored their systems not only to meet the task's technical challenges but also to address its social and linguistic intricacies. The strong use of domain knowledge whether in the form of handcrafted features, curated lexicons, or transliteration-informed pipelines proved essential in bridging the gap between model generalization and task-specific sensitivity. These contributions underscore a grow-

ing maturity in how the NLP community is approaching misinformation detection, especially in multilingual, code-mixed, and ethically charged contexts.

6 Result and Discussion

There were 11 participating teams who applied various approaches to tackle racial hoax detection in noisy Hindi-English code-mixed social media comments. Table 2 displays the final rank list for our shared task. We identified to rank the teams basis the Macro F1 evaluation metric. The scores are displayed are in the descending order of macro F1 score. The Team CUET's White_Walkers' staged fine-tuning of XLM-R achieved Rank 1 with goog implementation of training techniques. The Hope_for_best stability-focused MuRIL model with preprocessing texts with stratified class balanced training. The Team Dll5143A second run used language tags at the word level followed by character level embeddings and stereotype term dictionary in the third run providing new set of linguistic features. The Team KCRL followed the ensemble method from XLM-RoBERTa, MuRIL,and DeBERTa which was unique approach. The same ensemble method was followed by Team Hoax Terminators with their first run a combination of MuRIL and XLM-R. Most teams leveraged multilingual transformer models XLM-RoBERTa, MuRIL, mBERT, DeBERTa, with preprocessing pipelines addressing code-switching complexities through transliteration, language tagging, and noise reduction. The team EM-26 adopted a methodology that integrated traditional fine-tuning with lexicon-aware normalization. Their approach specifically addressed inconsistencies in the spelling and transliteration of Hindi words by employing normalization strategies and token filtering during the preparation of training data.

7 Conclusion

We presented the first ever shared task on detecting racial hoaxes in code-mixed Hindi-english social media data. We expect this task to have a great influence on low resource languages especially in the NLP domain because we received a variety of submissions with various methodologies. The most successful submissions applied were model finetuning with good training strategies, ensemble based methods and addressing lexicon aware normalization for code mixed Hindi-English words. The prediction evaluation was evaluated with a macro F1 score. The task showcased improvement in training techniques leveraging domain knowledge through tailored features and informed processing approaches has been key to enhancing model performance on specialized tasks reducing the gap between model generalization and task specific language sensitivity.

Acknowledgments

Author Bharathi Raja Chakravarthi was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight). This work is also supported by the Centre for Research Training in Artificial Intelligence grant number SFI/18/CRT/6223, as well as a grant from the College of Science and Engineering, University of Galway, Ireland.

References

Tewodros Achamaleh, Fatima Uroosa, Nida Hafeez, Abiola T. O., Mikiyas Mebiratu, Sara Getachew, Grigori Sidorov, and Rolando Quintero. 2025. Em-26@lt-edi 2025: Detecting racial hoaxes in code-mixed social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Cynthia Amol, Lilian Wanzare, and James Obuhuma. 2024. Politikweli: A Swahili-English code-switched Twitter political misinformation classification dataset. In *Speech and Language Technologies for Low-Resource Languages*, pages 3–17, Cham. Springer Nature Switzerland.

S Nagesh Bhattu, Satya Krishna Nunna, Durvasula VLN Somayajulu, and Binay Pradhan. 2020. Improving code-mixed POS tagging using code-mixed embeddings. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4):1–31.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An italian social media corpus where psychology meets nlp. *Information Processing & Management*, 60(1):103118.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696.

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, 18(1):49–68.

Arkadipta De, Dibyanayan Bandyopadhyay, Baban Gain, and Asif Ekbal. 2021. A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–20.

Anik Dey and Pascale Fung. 2014. A hindi-english code-switching corpus. In *LREC*, pages 2410–2413.

Arunava Kar and Angshuman Jana. 2024. Evaluating YouTube video via sentiment analysis: A case study in code-mixed Bangla-English context. In *Intelligent Systems Design and Applications*, pages 428–437, Cham. Springer Nature Switzerland.

Piyush Katariya, Vedika Gupta, Rohan Arora, Adarsh Kumar, Shreya Dhingra, Qin Xin, and Jude Hemanth. 2022. A deep neural network-based approach for fake news detection in regional language. *International Journal of Web Information Systems*, 18(5/6):286–309.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lori Kido Lopez. 2022. *Race and Digital Media: An Introduction*. John Wiley & Sons.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Abrar Hafiz Rabbani, Diganta Das Droba, Momtazul Arefin Labib, Samia Rahman, and Hasan Murad. 2025. Hoax terminators@lt-edi 2025: Charbert's dominance over llm models in the detection of racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Md. Abdur Rahman, MD AL AMIN, Sabik Aftahee, and Md Ashiqur Rahman. 2025a. Hinterwelt@lt-edi 2025: A transformer-based approach for identifying racial hoaxes in code-mixed hindi-english social media narratives. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Md Mizanur Rahman, Jidan Al Abrar, Md Siddikul Imam Kawser, Ariful Islam, Md. Mubasshir Naib, and Hasan Murad. 2025b. Cuet's_white_walkers@lt-edi 2025: Racial hoax detection in code-mixed on social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Arushi Sharma, Anubha Kabra, and Minni Jain. 2022. Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language. *Information Processing & Management*, 59(1):102760.

Rajalakshmi Sivanaiah, Nishaanth Ramanathan, Shajith Hameed, Rahul Rajagopalan, Angel Deborah Suseelan, and Mirnalinee Thanka Nadar Thanagathai. 2022. Fake news detection in low-resource languages. In *International conference on speech and language technologies for low-resource languages*, pages 324–331. Springer.

Abhishek Srivastava, Kalika Bali, and Monojit Choudhury. 2020. Understanding script-mixing: A case study of Hindi-English bilingual Twitter users. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 36–44, Marseille, France. European Language Resources Association.

Malliga Subramanian, Aruna A, Amudhavan M, Jahaganapathi S, and Kogilavani S V. 2025. Kec-elite-analysts@lt-edi 2025: Leveraging deep learning for racial hoax detection in code-mixed hindi-english tweets. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Gauri Takawane, Abhishek Phaltankar, Varad Patwardhan, Aryan Patil, Raviraj Joshi, and Mukta S. Takalikar. 2023. Language augmentation approach for code-mixed text classification. *Natural Language Processing Journal*, 5:100042.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus

creation and emotion prediction for Hindi-English code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Abhishek Singh Yadav, Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2025. Hope_for_best@lt-edi 2025: Detecting racial hoaxes in code-mixed hindi-english social media data using a multi-phase fine-tuning strategy. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Ashok Yadav and Vrijendra Singh. 2025. Dll5143a@lt-edi 2025: Bias-aware detection of racial hoaxes in code-mixed social media data (bacohoax). In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Overview of Fourth Shared Task on Homophobia and Transphobia Span Detection in Social Media Comments

Prasanna Kumar Kumaresan[1], Bharathi Raja Chakravarthi[1], Ruba Priyadharshini[2], Paul Buitelaar[3], Malliga Subramanian[4], Kishore Kumar Ponnusamy[5]

[1]School of Computer Science, University of Galway, Ireland
[2]Gandhigram Rural Institute – Deemed to be University, Tamil Nadu, India
[3]Insight SFI Research Centre for Data Analytics,
Data Science Institute, University of Galway, Ireland
[4]Kongu Engineering College, Tamil Nadu, India
[5]Digital University of Kerala, India

Correspondence: P.Kumaresan1@universityofgalway.ie

Abstract

The rise and the intensity of harassment and hate speech on social media platforms against LGBTQ+ communities is a growing concern. This work is an initiative to address this problem by conducting a shared task focused on the detection of homophobic and transphobic content in multilingual settings. The task comprises two subtasks: (1) multi-class classification of content into homophobia, transphobia, or non-anti-LGBT+ categories across eight languages and (2) span-level detection to identify specific toxic segments within comments in English, Tamil, and Marathi. This initiative helps the development of explainable and socially responsible AI tools for combating identity-based harm in digital spaces. Multiple teams registered for the task; however, only two teams submitted their results, and the results were evaluated using the macro F1 score.

1 Introduction

Homophobia and transphobia refer to harmful attitudes and prejudices directed toward individuals who identify as homosexual or transgender[1] (Hill, 2003; O'Donohue and Caselles, 1993; Nagoshi et al., 2008). While the terms may linguistically suggest irrational fear, they more accurately encompass a spectrum of negative biases and discriminatory behaviors against people who are lesbian, gay, bisexual, or transgender (Rollè et al., 2014). These biases can manifest in various forms, ranging from subtle expressions such as derogatory language to overt acts of hostility and aggression, contributing significantly to the marginalization and emotional distress experienced by LGBTQ+ individuals (Moagi et al., 2021).

Recently, the growth of social media has both amplified these challenges and created new avenues for their expression (Fuchs, 2014; Chakravarthi et al., 2022). While these platforms foster connection and community building, they have also become grounds for the spread of toxic language, including hate speech targeting LGBTQ+ communities (Kumaresan et al., 2023; Calderón et al., 2024). According to the European Union, 50% of LGBT persons have been victims of hate speech or hate crime [2]. Such homophobic and transphobic content online not only reinforces societal prejudices but also inflicts psychological harm (Newcomb and Mustanski, 2010). Therefore, the ability to detect and address such harmful language in social media content is essential for cultivating safer, more inclusive digital environments (Chakravarthi, 2024).

This shared task addresses the problem of homophobia and transphobia detection in social media comments. It aims to promote research into the automatic identification and classification of homophobic and transphobic language, with a particular focus on multilingual and under-resourced language contexts. The shared task comprises two components: comment-level classification (Kumaresan et al., 2023) and span-level detection (Kumaresan et al., 2025). This involves highlighting the exact phrases that serve as evidence for the classification, enabling a more fine-grained and interpretable analysis. Span detection is particularly valuable for building explainable NLP systems that not only flag harmful content but also provide trans-

[1] https://reportandsupport.qmul.ac.uk/support/what-is-homophobia-transphobia-acephobia-and-biphobia

[2] https://fra.europa.eu/sites/default/files/fra_uploads/1226-Factsheet-homophobia-hate-speech-crime_EN.pdf

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 228–233

parent justifications for their decisions (Naim et al., 2022).

The dataset used for this task is derived from the manually annotated homophobia/transphobia content, which includes YouTube comments labeled at the comment level. Participants are encouraged to develop robust NLP systems capable of accurately identifying and categorizing hate speech targeting LGBTQ+ individuals. By tackling both classification and span detection, this shared task provides a platform for the NLP community to advance techniques for harmful content detection while fostering socially responsible NLP research across diverse linguistic and cultural settings.

In the upcoming section, we will describe the task description, dataset statistics, and participants' methodology towards the investigation of homophobia and transphobia detection from the YouTube comments on Dravidian languages.

2 Related Works

Span detection, also known as span-based classification or span identification, involves pinpointing the specific segments of a text that contain harmful or toxic content, rather than labeling the entire text as toxic (Pavlopoulos et al., 2021). This approach is particularly valuable in scenarios where only a small portion of a comment or post contains offensive language, while the remainder may be benign or contextually neutral (Gu et al., 2022). Traditional text classification models typically assign a single label to the entire input, which can be limiting in practical content moderation settings. Flagging an entire message as toxic based on a minor fragment may lead to unnecessary censorship and hinder constructive discourse.

Recent research in hate speech detection has increasingly emphasized the importance of explainability and precision (Sawant et al., 2024; Calabrese et al., 2024). Span-level annotations offer moderators actionable insights by highlighting the exact portions of the text that violate community guidelines, thereby streamlining the moderation process and enabling more targeted interventions (Mathew et al., 2021). This is especially crucial in social media contexts where high volumes of user-generated content make manual review inefficient.

In the context of homophobia and transphobia, span detection plays a critical role in identifying instances of identity-based harm (Zhou et al., 2023). Recent studies such as (Kumaresan et al., 2024) have explored the use of fine-grained annotations to detect hate speech against LGBTQ+ individuals, highlighting the need for datasets and models that capture identity-specific slurs and implicit hate spans. Studies (Condom Tibau et al., 2025; Chakravarthi et al., 2024) further illustrate the challenges in reliably detecting toxic content targeted at LGBT communities, showing that span-based approaches can improve both precision and fairness in these cases. These advances underscore the value of targeted span detection for moderating homophobic and transphobic content, offering more transparent and inclusive systems for content moderation.

Languages	Set	H	T	N
English	Train	179	7	2,978
	Dev	42	2	748
	Test	55	4	931
Tamil	Train	453	145	2,064
	Dev	118	41	507
	Test	152	47	634
Malayalam	Train	476	170	2,468
	Dev	197	79	937
	Test	140	52	674
Telugu	Train	2,907	2,647	3,496
	Dev	588	605	747
	Test	624	571	744
Kannada	Train	2,765	2,835	4,463
	Dev	585	617	955
	Test	599	606	951
Gujarati	Train	2,267	2,004	3,848
	Dev	498	454	788
	Test	510	436	794
Hindi	Train	45	92	2,423
	Dev	2	13	305
	Test	3	10	308
Marathi	Train	551	377	2,572
	Dev	129	80	541
	Test	112	69	569

Table 1: Multilingual classification (Task 1) dataset statistics (H-Homophobia, T-Transphobia, and N-Non-anti-LGBT+ content)

3 Task Description

We organized the shared task on homophobia & transphobia with around two subtasks.

- *Subtask 1*: Homophobia & Transphobia Multilingual Classification Task

- Objective: Classify comments into three categories: Homophobia, Transphobia, and None of the Above.
- Languages: This task will be conducted in multiple languages, specifically English, Tamil, Malayalam, Hindi, Gujarati, Telugu, Kannada, and Marathi.
- Special Focus on Tulu: Given the scarcity of resources like annotated corpora for under-resourced languages like Tulu, this task presents a unique challenge. We have introduced a code-mixed Tulu dataset specifically designed to detect homophobic and transphobic content. This dataset aims to promote research in few-shot learning, pushing the boundaries of what's possible in language processing for low-resource contexts.

- *Subtask 2*: Homophobia & Transphobia Span Detection
  - Objective: Identify specific spans within comments that contain instances of homophobia and transphobia.
  - Languages: English, Tamil, and Marathi.
  - Details: Participants will be provided with comments and are required to classify these comments at the span level. This task requires a deeper level of text understanding and precision, as participants must discern and highlight the textual evidence for homophobia or transphobia within the comments.

Overall, these tasks are designed not only to address significant technical challenges in the field of NLP but also to contribute to social good by identifying and mitigating harmful content directed at the LGBTQ+ community in diverse linguistic contexts.

4 Dataset Statistics

Social media platforms Twitter, Facebook, and YouTube use user-generated content to shape public opinion, which affects how people perceive things and how they view others. Recognizing the growing need for automated tools to extract emotions and detect harmful or irrelevant content online, particularly on platforms like YouTube, where user comments are rapidly increasing, we focused on content relevant to the LGBTQ+ community, who frequently engage with such platforms to express their views on various topics.

Languages	Set	H	T	N
Tamil	Train	188	75	137
	Test	73	36	63
English	Train	117	39	44
	Test	49	17	20
Marathi	Train	253	119	123
	Test	108	53	52

Table 2: Span Detection (Task 2) dataset statistics (H-Homophobia, T-Transphobia, and N-None of above)

We gathered a multilingual collection of YouTube comments about LGBTQ+ for Task 1. We protected individual privacy by not including personal stories from LGBTQ+ individuals in our collection. Using the YouTube Comment Scraper tool, we collected comments and manually annotated them with one of three labels: homophobic, transphobic, and non-anti-LGBT+ content. The final dataset languages - English, Tamil, Malayalam, Telugu, Kannada, Gujarati, Hindi, and Marathi were annotated following the guidelines outlined in the dataset paper (Kumaresan et al., 2023). The distribution of annotated labels across all languages appears in Table 1.

For Task 2, we extended our efforts by annotating spans of text within comments that explicitly or implicitly expressed homophobia or transphobia (Kumaresan et al., 2025). These span-level annotations were carried out in three languages, Tamil, English, and Marathi, using the sequence labeling approach implemented in the open-source annotation tool Doccano. We focused on marking only those portions of text that conveyed discriminatory attitudes, allowing us to take a targeted and strategic annotation approach. Table 2 shows the dataset statistics for span annotations across the three categories: Homophobia (H), Transphobia (T), and Non-anti-LGBT+ content (N).

5 Participants Methodology

We organized a shared task focused on addressing harmful content that targets LGBTQ+ individuals through two essential subtasks. The participants used multiple machine learning and deep learning approaches to tackle these subtasks, especially when working with low-resource and multilingual data.

The *SKV TRIO team* (Vignesh et al., 2025) used a combination of BERT and TF-IDF embeddings for Task 1. The team used BERT and TF-IDF embeddings for each input before applying dimensionality reduction to TF-IDF embeddings to match BERT's dimensions. The system combined these embeddings to create a single feature representation, which served as input for training a random forest classifier. The method united semantic depth with statistical feature patterns to produce an interpretable and efficient computational solution.

The *KEC-Elite-Analysts* team used multiple deep learning models to solve task 1 by classifying homophobia and transphobia. The architecture used bidirectional LSTM and GRU models to extract sequential and contextual language patterns and class weights to handle class imbalance. A TextCNN module to detect local n-gram features indicative of toxic expressions. A multilayer perceptron (MLP) trained on averaged word embeddings to incorporate semantic information into the final prediction.

The models were designed to generalize across multiple languages, including low-resource and code-mixed languages such as Tamil. This multilingual focus ensured robust performance in linguistically diverse and underrepresented languages.

6 Result and Discussion

A total of 30 participants registered for our shared task. Nevertheless, only two teams submitted results for Task 1, and no submissions were received for Task 2. The fact that Task 2 required identifying specific spans of homophobic and transphobic content may have contributed to its lack of submissions. This probably required more domain knowledge and work, which might have made it difficult for participants to finish in the allotted time.

The outcomes of *Task 1* are displayed. The macro F1 scores for the two participating teams, SKV TRIO and KEC-Elite-Analysts, across the supported languages are shown in Table 3. To take into consideration the dataset's multilingual nature and multi label task, the evaluation was carried out independently for each language with macro F1 score. Because it computes the F1 score for each class separately and then averages them, treating all classes equally regardless of size, we decided to use the macro F1 score to assess the ranklist result.

In the majority of languages, including Gujarati (0.86), Telugu (0.87), and Kannada (0.81), the *SKV TRIO* team received the highest scores. Their method of training a random forest classifier by combining BERT and TF-IDF embeddings seems to have successfully identified both statistical and semantic patterns in the data. Low-resource and morphologically rich languages may have benefited most from this hybrid embedding approach, as term-level distinctions unique to hate speech patterns are reinforced by TF-IDF, while pre-trained contextual models such as BERT can offer general language understanding. The model's strong performance in languages with limited resources and varying the dimensionality alignment between embeddings, which also helped the model generalize better across a variety of linguistic structures.

The *KEC-Elite-Analysts* team outperformed the SKV TRIO team in English (0.40) and Tamil (0.74), demonstrating notable competence in those languages. Their system used an MLP trained on averaged word embeddings in conjunction with a deep learning ensemble comprising Bidirectional LSTM, GRU, and TextCNN components. This architecture works well with languages like English, where pre-trained embeddings and deep learning models typically perform reliably due to an abundance of resources, and Tamil, where code-mixing and sequential dependencies are common. Their system was able to capture subtle patterns in sentence structure, particularly in high-resource or semi-structured languages, because of the ensemble design and the use of class weights to address label imbalance.

These findings show that various modeling approaches obtained performance differences across languages, which may have been caused by the variety of languages and the accessibility of data. While KEC-Elite-Analysts' deep learning ensemble approach proved successful in identifying patterns in more resource-intensive or frequently used languages like English and Tamil, SKV TRIO's fusion-based feature engineering demonstrated superior generalization across a wider range of languages. The findings highlight how crucial model diversity and adaptability are, especially when working in environments with limited resources and code-mixed languages. They also highlight the need for more research into span-level detection, since future versions of the task might benefit from longer timeframes, more annotation support, or easier baselines for span identification to reduce the barrier to entry.

Team Name	Task 1: Languages - Macro F1 Score							
	English	Gujarathi	Hindi	Tamil	Telugu	Marathi	Malayalam	Kannada
SKV TRIO (Vignesh et al., 2025)	0.34	**0.86**	**0.33**	0.37	**0.87**	0.29	**0.40**	**0.81**
KEC-Elite-Analysts (Run 1)	**0.40**	-	-	**0.74**	-	**0.52**	-	-

Table 3: Results from Task 1 showing macro F1 scores by language for each participating team (bold values indicate the highest score per language).

7 Conclusion

In this shared task, we addressed the challenge of two sub-tasks, which are detecting homophobia and transphobia classification and span identification through multilingual and low-resource languages. A total of 30 participants were registered, only two teams submitted the results for Task 1, and no submissions were received for Task 2, likely due to the complexity of span annotation and the need for domain-specific understanding within a limited timeframe. The classification results showed the efficacy of various modeling approaches, with deep learning ensembles performing well in high-resource languages and hybrid embedding approaches excelling in low-resource contexts. These results emphasize how crucial flexible, language-sensitive models are for identifying harmful content. Although span detection remains a challenging and underexplored area, specifically in low-resource, it is critical for the development of explainable and culturally aware moderation systems. Future iterations of this task should aim to reduce entry barriers and further promote research in inclusive and socially responsible NLP.

Acknowledgments

Author Bharathi Raja Chakravarthi was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight). This work is also supported by the Centre for Research Training in Artificial Intelligence grant number 18/CRT/6223, and a grant from the College of Science and Engineering, University of Galway, Ireland.

References

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 398–408, Bangkok, Thailand. Association for Computational Linguistics.

Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. 2024. From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and lgbt communities. *Humanities and Social Sciences Communications*, 11(1):1369. Published: October 15, 2024.

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, 18:49–68.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Jordi Guillem Condom Tibau, Angelina Voggenreiter, elena pavan, and Jürgen Pfeffer. 2025. Prevalence, substance and responses to hate speech against lgbtq communities on tiktok. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):430–442.

Christian Fuchs. 2014. *Social Media: A Critical Introduction*. SAGE Publications Ltd, London.

Weiwei Gu, Boyuan Zheng, Yunmo Chen, Tongfei Chen, and Benjamin Van Durme. 2022. An empirical study on finding spans. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3976–3983, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Darryl B Hill. 2003. Genderism, transphobia, and gender bashing: A framework for interpreting anti-transgender violence. In *Understanding and dealing with violence: A multicultural approach*, pages 113–136. SAGE Publications, Inc.

Prasanna Kumar Kumaresan, Devendra Deepak Kayande, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2025. Homophobia and transphobia span identification in low-resource languages. *Natural Language Processing Journal*, 12:100169.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, 5:100041.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. Dataset for identification of homophobia and transphobia for Telugu, Kannada, and Gujarati. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411, Torino, Italia. ELRA and ICCL.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

M.M. Moagi, A.E. van Der Wath, P.M. Jiyane, and R.S. Rikhotso. 2021. Mental health challenges of lesbian, gay, bisexual and transgender people: An integrated literature review. *Health SA Gesondheid*, 26:1487.

Julie L Nagoshi, Katherine A Adams, Heather K Terrell, Eric D Hill, Stephanie Brzuzy, and Craig T Nagoshi. 2008. Gender differences in correlates of homophobia and transphobia. *Sex roles*, 59:521–531.

Jannatun Naim, Tashin Hossain, Fareen Tasneem, Abu Nowshed Chy, and Masaki Aono. 2022. Leveraging fusion of sequence tagging models for toxic spans detection. *Neurocomputing*, 500:688–702.

Michael E. Newcomb and Brian Mustanski. 2010. Internalized homophobia and internalizing mental health problems: A meta-analytic review. *Clinical Psychology Review*, 30(8):1019–1029.

William O'Donohue and Christine E Caselles. 1993. Homophobia: Conceptual, definitional, and value issues. *Journal of Psychopathology and Behavioral Assessment*, 15:177–195.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Luca Rollè, Piera Brustia, and Angela Caldarera. 2014. *Homophobia and Transphobia*, pages 2905–2910. Springer Netherlands, Dordrecht.

Madhuri Sawant, Arjumand Younus, Simon Caton, and Muhammad Atif Qureshi. 2024. Using explainable ai (xai) for identification of subjectivity in hate speech annotations for low-resource languages. In *Proceedings of the 4th International Workshop on Open Challenges in Online Social Networks*, OASIS '24, page 10–17, New York, NY, USA. Association for Computing Machinery.

Konkimalla Laxmi Vignesh, Mahankali Sri Ram Krishna, Dondluru Keerthana, and Premjith B. 2025. Skv trio@lt-edi-2025: Hybrid tf-idf and bert embeddings for multilingual homophobia and transphobia detection in social media comments. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. 2023. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, 29(5):1247–1274.

Overview of the Fifth Shared Task on Speech Recognition for Vulnerable Individuals in Tamil

B. Bharathi[1], **Bharathi Raja Chakravarthi**[2],
N. Sripriya[1], **Rajeswari Natarajan**[3], **Rajalakshmi R** [4], **S. Suhasini**[5]
[1]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
[2]School of Computer Science, University of Galway, Ireland
[3]SASTRA University, India
[4]VIT Chennai, India
[5]Saveetha Engineering College, Tamil Nadu India
bharathib@ssn.edu.in, bharathiraja.akr@gmail.com

Abstract

In this paper, an overview of the shared task on speech recognition for vulnerable individuals in Tamil (LT-EDI@LDK2025) is described. The work comes with a Tamil dataset that was collected from elderly individuals who identify as male, female, or transgender. The audio samples were taken in public places such as markets, vegetable shops, hospitals, etc. The training phase and the testing phase are when the dataset is made available. The task required of the participants was to handle audio signals using various models and techniques and then turn in their results as transcriptions of the provided test samples. The participant's results were assessed using WER (Word Error Rate). The transformer-based approach was used by participants to achieve automatic voice recognition. This overview paper discusses the findings and various pre-trained transformer-based models that the participants employed.

1 Introduction

The earliest known examples of Old Tamil writing are tiny inscriptions found in Adichanallur that date between 905 and 696 BC. Of all the Indian languages, Tamil possesses the most ancient non-Sanskritic literature. The grammar of Tamil is agglutinative, meaning that noun class, number, case, verb tense, and other grammatical categories are indicated by suffixes. Unlike other Aryan languages, which use Sanskrit as their standard language, Tamil uses Tamil for both its scholarly vocabulary and its metalinguistic terminology. Together with dialects, Tamil has multiple forms: cankattami, the classical literary style based on the ancient language; centami, the modern literary and formal style; and kotuntami, the present vernacular form. (Sakuntharaj and Mahesan, 2021, 2017). There is a stylistic continuity created by these styles merging together. For instance, one may write centami using cankattami vocabulary, or one could speak kotuntami while using forms related to one of the other types. (Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). A lexical root plus one or more affixes combine to form Tamil words. Suffixes make up the bulk of affixes in Tamil. Tamil suffixes fall into two groups: derivational suffixes, which change a word's meaning or part of speech, and inflectional suffixes, which identify certain categories like person, number, mood, tense, and so on. Agglutination can lead to huge words with multiple suffixes, needing numerous words or a phrase in English. Its length and scope are infinite. Although smart technologies have come a long way, human-machine interaction is still being developed and enhanced. (Chakravarthi et al., 2020). Automatic speech recognition (ASR) is one such recent technology that has enabled voice-based user interfaces for numerous automated systems. Many elderly and transgender people are frequently unaware of the technology (Hämäläinen et al., 2015) that is made available to help people in public places like banks, hospitals, and administrative offices. Thus, communication is the only kind of media that can assist people in getting what they want. However these ASR systems are infrequently used by the elderly, transsexuals, and others with lower levels of education. English-language voice-based interfaces are a feature of most automated systems currently in use. Elderly people and those living in rural areas prefer to speak in their native tongue. The provision of speech interfaces in the local language for help systems designed for public usage would be advantageous to all. Information regarding spontaneous speech in Tamil is gathered from transgender and elderly people who are not able to use these programs. The aim of this challenge is to find an efficient ASR model to handle the elderly person's speech corpus. The representation of how the audio samples are collected is shown in Fig:1

The pertinent features will first be extracted from the speech signal using an ASR system. Acoustic

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 234–240

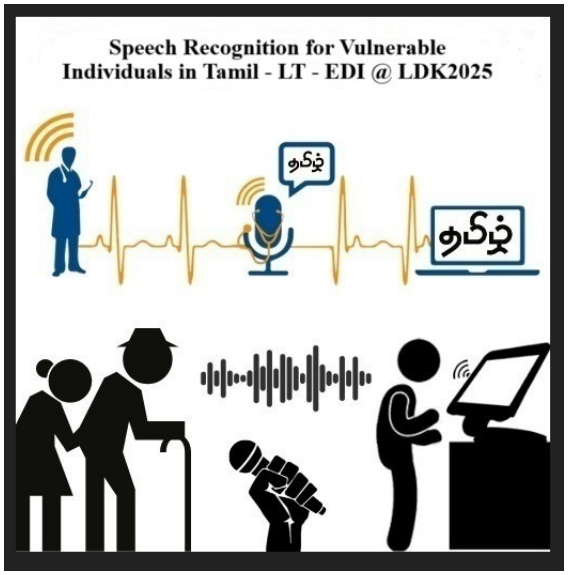

Figure 1: Speech corpus collected from vulnerable individuals in Tamil language

models will also be produced using these features that were retrieved. Ultimately, the language model assists in converting these probabilities into grammatical words. The language model uses statistics from training data to assign probabilities to words and phrases (Das et al., 2011). It is necessary to evaluate ASR systems' performance prior to deploying them in real-time applications. On large-scale automatic speech recognition (ASR) tasks, an end-to-end speech recognition system has shown promising performance, matching or surpassing that of traditional hybrid systems. Using an acoustic model, lexicon, and language model, the end-to-end system quickly transforms audio data into tag labels (Zeng et al., 2021; Pérez-Espinosa et al., 2017). In the field of end-to-end voice recognition, there exist two extensively utilized frameworks. Frame synchronous prediction separates one input frame from the other by giving each one a target label (Miao et al., 2020, 2019). Phoneme identification can also be used to assess the efficacy using different test feature vectors and model settings. The use of acoustic models for speech recognition, which are created using the sounds of younger people, may have a substantial impact on the capacity to recognize elder speech (Fukuda et al., 2020; Zeng et al., 2020). There aren't many acoustic models that can handle the voice detection task. Among the acoustic models are Japanese Newspaper Article Sentences (JNAS), Japanese Newspaper Article Sentences Read Speech Corpus of the Aged (S-JNAS), and Corpus of Spontaneous Japanes (CSJ). The CSJ model only achieves the lowest WER once the older voices are adjusted, according to a comparison of all the acoustic models in the literature (Fukuda et al., 2020). Dialect adaptation is also required in order to improve recognition accuracy (Fukuda et al., 2019). Recent advances in large vocabulary continuous speech recognition (LVCSR) technologies have led to the widespread use of speech recognition systems in several fields (Xue et al., 2021). Variations in the acoustics of individual speakers are thought to be one of the primary causes of the decline in speech recognition rates. For elder speakers to use speech recognition systems trained on typical adult speech data, the acoustic discrepancies between their speech and that of an adult should be investigated and correctly adjusted. Rather, this loss can be mitigated by an acoustic model enhanced by senior speakers' utterances, as shown by a document retrieval system. Modern voice recognition technology can reach excellent recognition accuracy while speaking while reading a written text or something comparable; nevertheless, the accuracy decreases when speaking spontaneously and freely. The main reason for this issue is that the linguistic and acoustic models used in voice recognition were mostly developed using read aloud or written language materials. However, there are significant linguistic and auditory differences between written language and spontaneous speech(Zeng et al., 2020). Currently, it is becoming more and more popular to create ASR systems that can detect voice data from older persons. The aging population in modern society and the proliferation of smart devices, which make information freely accessible to both the young and the old, have led to a demand for improved voice recognition in smart devices (Kwon et al., 2016; Vacher et al., 2015; Hossain et al., 2017). Because of the influences of speech articulation and speaking style, speech recognition systems are often optimized for the voice of an average adult and have a lower accuracy rate when recognising the voice of an elderly person. It will surely become more expensive to adapt the current voice recognition systems to handle the speech of elderly users(Kwon et al., 2016).

2 Task Description

This shared task tackles a difficult problem in Automatic Speech Recognition: vulnerable elderly and transgender individuals in Tamil. People in their se-

nior years go to primary places such as banks, hospitals, and administrative offices to meet their daily needs. Many elderly persons are unsure of how to use the devices provided to assist them. Similarly, because transgender persons are denied access to primary education as a result of societal discrimination, speech is the only channel via which they may meet their needs. The data on spontaneous speech is collected from elderly and transgender people who are unable to take advantage of these services. For the training set, a speech corpus containing 5.5 hours of transcribed speech will be released, as well as 2 hours of speech data for testing test. The participants have to submit the text transcriptions for the test utterances in a separate text file.

3 Related Work

When a model is fine-tuned on many languages at the same time, a single multilingual speech recognition model can be built that can compete with models that are fine-tuned on individual language speech corpus. Speech2Vec expands the text-based Word2Vec model to learn word embeddings directly from speech by combining an RNN Encoder-Decoder framework with skipgrams or cbow for training. Acoustic models are designed at phoneme/syllable level to carry out the speech recognition task. Initially, the acoustic models were created with JNAS, S-JNAS and CSJ speech corpus(Lin and Yu, 2015; Iribe et al., 2015). Later, the models were trained/fine-tuned with different speech corpus. To get a better performance and accuracy, backpropagation using the transfer learning was attempted in the literature. Similar work was performed for other languages like Bengali, Japanese, etc. Also, more speech corpus is collected from the young people for many languages(Zeng et al., 2020; Lee et al., 2021). However, speaker fluctuation, environmental noise, and transmission channel noise all degrade ASR performance. As the shared task is given with a separate training data set, an effective model has to be created during the training. Therefore, hierarchical transformer based model for large context end to end ASR can be used (Masumura et al., 2021). In the recent era, the environment is changing with smart systems and is identified that there is a need for ASR systems that are capable of handling speech of elderly people spoken in their native languages. To overcome this problem, the shared task is proposed for the research community to build an efficient model for recognizing the speech of elderly people and transgenders in Tamil language. Findings of the automatic speech recognition for vulnerable individuals are given in (S and B, 2022) (B et al., 2022)("S and B, "2023") (Bharathi et al., 2023), have used transformer models used for transformer based ASR for Vulnerable Individuals in Tamil.

4 Data-set Description

The dataset given to this shared task (Bharathi et al., 2022) is an Tamil conversational speech recorded from the elderly people whose average age is around 61 for male, 59 for female and 30 for transgender people. A total of 7.5 hours is collected from the elderly people. 46 audio files were recorded and each audio file is split into many subsets as transformer model does not support the large audio files. The speech is recorded with a sampling rate of 16KHZ. The audio files from Audio - 1 to Audio - 36 are used for training (duration is approximately 5.5 hours) and Audio - 37 to Audio - 51 are used for testing (duration is approximately 2 hours).

5 Methodology

The methodology used by the participants in shared task of speech recognition for vulnerable individuals in Tamil is discussed in this section.

- **NSR:** The Team fine-tuned OpenAI's Whisper v3 Large model for Tamil speech recognition using the Common Voice Tamil dataset. The dataset was preprocessed by normalizing transcripts and ensuring alignment between audio files and text. We employed Connectionist Temporal Classification (CTC) Loss for training, optimizing the model using the AdamW optimizer with a learning rate of 3e-4 and a warm-up phase. We used gradient accumulation to handle large batch sizes efficiently and mixed precision training (FP16) for faster convergence. Evaluation was conducted using Word Error Rate (WER) and Character Error Rate (CER). The fine-tuning was performed using the Hugging Face Transformers library, leveraging PyTorch and GPU acceleration to speed up training.

- **CrewX:** The speech signal was initially denoised using Adaptive Variational Mode Decomposition (AVMD), which decomposes it

into variational modes and reconstructs the relevant components to remove noise while preserving speech clarity. Next, Silero VAD (GRU based) was applied to eliminate silence and non-speech segments, reducing computational load due to unnecessary processing and at the same time improving transcription accuracy.The processed audio is then passed through the Whisper processor, which converts it into log-Mel spectrogram features using a standardized pipeline. Finally, the extracted features are passed to the Whisper-Tamil- Medium model for generating transcriptions, with beam search decoding during testing to enhance the accuracy and reduce the WER.

- **JUNLP:** The speech recognition was performed using two pre-trained state-of-the-art models, Whisper and XLSR. Both models were trained on the Tamil corpus. Whisper is a pre-trained automatic speech recognition (ASR) model trained on 680,000 hours of multilingual and multitask supervised data sourced from the web. They utilized vasista22/whisper-tamil-large-v21 model which is fine-tuned version of openai/whisper-large-v22 on the Tamil data available from multiple publicly available ASR corpuses. This transformer-based encoderdecoder model processes log-Mel spectrograms through convolutional layers in the encoder and generates text autoregressively in the decoder. The model was further fine-tuned on a Tamil corpus of given training dataset, providing a robust baseline for Tamil speech recognition. To adapt the 1.59-billion-parameter Whisper model efficiently, we utilize Low-Rank Adaptation (LoRA)and Dynamic Rank Adaptation (DoRA). These techniques freeze pre-trained weights and inject trainable low-rank matrices into specific transformer submodules, reducing computational overhead while preserving model performance. On the other hand, They fine-tuned the pretrained anuragshas/wav2vec2-xlsr-53-tamil3 checkpoint with the Hugging Face Trainer API. The model is a Wav2Vec2ForCTC type model and was fine-tuned with full-scale finetuning, without layer freezing or modifications. Connectionist Temporal Classification (CTC) loss was used during training and performance was tracked with Word Error Rate (WER) and Character Error Rate (CER). Mixed precision training was activated with fp16=true, and the best model was chosen based on the minimum WER on the evaluation set. Gradient accumulation with an accumulation step of 2 was used to stabilize training and mimic larger batch sizes.

- **SSNCSE:** A fine-tuned version of OpenAI's Whisper model, known as yaygomii/FYP_Whisper_PEFT_TAMIL, is used as the base system. This model is adapted using Low-Rank Adaptation (LoRA), a technique under the Parameter-Efficient Fine-Tuning (PEFT) framework. LoRA allows a model to learn domain-specific features by introducing trainable low-rank matrices into pre-existing attention layers while freezing all but a small number of parameters. This reduces the computational burden and memory use in the training phase, making it a great option for low-resource scenarios.

6 Evaluation of Results

The results submitted by the participants are evaluated based on the WER computed between the ASR hypotheses submitted by the participants and the ground truth of human speech transcription.

$$\text{WER (Word Error Rate)} = (S + D + I) / N$$

where,
S = No. of substitutions
D = No. of deletions
I = No. of insertions
N = No. of words in the reference transcription

As discussed in the methodology, different average word error rate are measured using various pre-trained transformer based models.

7 Conclusions

The shared challenge for vulnerable voice recognition in Tamil is covered in this overview paper. The speech corpus shared for this job was recorded from elderly persons. Getting older people's speech more accurately recognised is a difficult endeavour. In order to boost the accuracy and performance in recognising the elderly people's speech, the participants have been given access to the gathered

S. No	Team Name	WER (in %)
1	NSR(S et al., 2025)	34.854
2	CrewX(Sundhar S et al., 2025)	31.897
3	JUNLP(Acharya et al., 2025)	38.428
4	SSNCSE(K and B, 2025)	42.306

Table 1: Results of the participating systems in Word Error Rate

speech corpus. There were totally six teams participated in this joint task and turned in their transcripts of the supplied data. The team estimated the WER and then compared the outcome to the human transcripts. Five teams built their recognition systems using various Whisper model and transformer-based models. Finally, the word error rates of the five participants are 34.854, 34.925, 31.897, 38.428, 42.306 respectively. Based on the observations, it is suggested that the transformer based model and whisper model can be trained with given speech corpus which could give a better accuracy than the pre-trained model, as the transformer based model and whisper model used are trained with common voice dataset. Also, a separate language model can also be created for this corpus.

References

Priyobroto Acharya, Soham Chaudhuri, Sayan Das, Dipanjan Saha, and Dipankar Das. 2025. Junlp@lt-edi-2025: Efficient low-rank adaptation of whisper for inclusive tamil speech recognition targeting vulnerable populations. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Bharathi B, Dhanya Srinivasan, Josephine Varsha, Thenmozhi Durairaj, and Senthil Kumar B. 2022. SSNCSE_NLP@LT-EDI-ACL2022:hope speech detection for equality, diversity and inclusion using sentence transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 218–222, Dublin, Ireland. Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Cn, Sripriya Natarajan, Rajeswari Natarajan, S Suhasini, and Swetha Valli. 2023. Overview of the second shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 31–37.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020. Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

Biswajit Das, Sandipan Mandal, and Pabitra Mitra. 2011. Bengali speech corpus for continuous auutomatic speech recognition system. In *2011 International conference on speech database and assessments (Oriental COCOSDA)*, pages 51–55. IEEE.

Meiko Fukuda, Ryota Nishimura, Hiromitsu Nishizaki, Yurie Iribe, and Norihide Kitaoka. 2019. A new corpus of elderly japanese speech for acoustic modeling, and a preliminary investigation of dialect-dependent speech recognition. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Meiko Fukuda, Hiromitsu Nishizaki, Yurie Iribe, Ryota Nishimura, and Norihide Kitaoka. 2020. Improving speech recognition for the elderly: A new corpus of elderly japanese speech and investigation of acoustic modeling for speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6578–6585.

Annika Hämäläinen, António Teixeira, Nuno Almeida, Hugo Meinedo, Tibor Fegyó, and Miguel Sales Dias. 2015. Multilingual speech recognition for the elderly: The aalfred personal life assistant. *Procedia Computer Science*, 67:283–292.

M Shamim Hossain, Md Abdur Rahman, and Ghulam Muhammad. 2017. Cyber–physical cloud-oriented multi-sensory smart home framework for elderly people: An energy efficiency perspective. *Journal of Parallel and Distributed Computing*, 103:11–21.

Yurie Iribe, Norihide Kitaoka, and Shuhei Segawa. 2015. Development of new speech corpus for elderly japanese speech recognition. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language*

Research and Evaluation (O-COCOSDA/CASLRE), pages 27–31. IEEE.

Sreeja K and Bharathi B. 2025. Ssncse@lt-edi-2025:speech recognition for vulnerable individuals in tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Soonil Kwon, Sung-Jae Kim, and Joon Yeon Choeh. 2016. Preprocessing for elderly speech recognition of smart devices. *Computer Speech & Language*, 36:110–121.

Taewoo Lee, Min-Joong Lee, Tae Gyoon Kang, Seokyeoung Jung, Minseok Kwon, Yeona Hong, Jungin Lee, Kyoung-Gu Woo, Ho-Gyeong Kim, Jiseung Jeong, and 1 others. 2021. Adaptable multi-domain language model for transformer asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7358–7362. IEEE.

Hui Lin and Yibiao Yu. 2015. Acoustic feature analysis and conversion of age speech. In *IET Conference Proceedings*. The Institution of Engineering & Technology.

Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi. 2021. Hierarchical transformer-based large-context end-to-end asr with large-context knowledge distillation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5879–5883. IEEE.

Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, and Yonghong Yan. 2020. Transformer-based online ctc/attention end-to-end speech recognition architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE.

Haoran Miao, Gaofeng Cheng, Pengyuan Zhang, Ta Li, and Yonghong Yan. 2019. Online hybrid ctc/attention architecture for end-to-end speech recognition. In *Interspeech*, pages 2623–2627.

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, and Himer Avila-George. 2017. Using acoustic paralinguistic information to assess the interaction quality in speech-based systems for elderly users. *International Journal of Human-Computer Studies*, 98:1–13.

Nishanth S, Shruthi Rengarajan, Burugu Rahul, and G. Jyothish Lal. 2025. Nsr@lt-edi-2025: Automatic speech recognition in tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Suhasini S and Bharathi B. 2022. SUH_ASR@LT-EDI-ACL2022: Transformer based approach for speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 177–182, Dublin, Ireland. Association for Computational Linguistics.

Suhasini "S and Bharathi" B. "2023". "asr_ssn_cse 2023@lt-edi-2023: Pretrained transformer based automatic speech recognition system for elderly people". In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, "Varna, Bulgaria ". "Recent Advances in Natural Language Processing".

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

Ganesh Sundhar S, Hari Krishnan N, Arun Prasad TD, Shruthikaa V, and Jyothish Lal G. 2025. Crewx@lt-edi-2025: Transformer-based tamil asr fine-tuning with avmd denoising and gru-vad for enhanced transcription accuracy. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Michel Vacher, Frédéric Aman, Solange Rossato, and François Portet. 2015. Development of automatic speech recognition techniques for elderly home support: Applications and challenges. In *International Conference on Human Aspects of IT for the Aged Population*, pages 341–353. Springer.

Jiabin Xue, Tieran Zheng, and Jiqing Han. 2021. Exploring attention mechanisms based on summary information for end-to-end automatic speech recognition. *Neurocomputing*, 465:514–524.

Jiazhong Zeng, Jianxin Peng, and Yuezhe Zhao. 2020. Comparison of speech intelligibility of elderly aged 60–69 years and young adults in the noisy and reverberant environment. *Applied Acoustics*, 159:107096.

Zhiping Zeng, Haihua Xu, Yerbolat Khassanov, Eng Siong Chng, Chongjia Ni, Bin Ma, and 1 others. 2021. Leveraging text data using hybrid transformer-lstm based end-to-end asr in transfer learning. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

Author Index