Audio Driven Detection of Hate Speech in Telugu: Toward Ethical and Secure CPS

Santhosh Kumar M[1*], Sai Ravula P[1*], Prasanna Teja M[1], Ajay Surya J[1], Mohitha V[1], Jyothish Lal G[1†]

[1*]School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, 641112, Tamilnadu, India.

*Corresponding author(s). E-mail(s): madha.santhoshkr@gmail.com; sairavula07@gmail.com;
Contributing authors: tejamuraharirao45@gmail.com; jaysatya11@gmail.com; mohithavelagapudi@gmail.com; g_jyothishlal@cb.amrita.edu;
†These authors contributed equally to this work.

Abstract

The rapid integration of social media platforms into cyber-physical systems has introduced new challenges in ensuring human-centric reliability and safety. This is mainly due to the widespread dissemination of hate speech and the inability of online systems to effectively moderate offensive content. While significant advances have been made toward hate speech detection in high-resource languages such as English, low-resource languages such as Telugu do not have the annotated datasets and tools to properly detect it. This project addresses this gap by creating a complete annotated multimodal hate speech dataset in the Telugu language, consisting of 2 hours of audio-text pairs from YouTube. The dataset enables the exploration of hate speech detection in individual modalities, speech and text, as well as in a combined multimodal setting. The work presented in this paper is focused on the detection of hate speech based on audio data with text-based analysis incorporated as an ablation study to better understand the modality-specific contributions. Our classification results demonstrate that the combination of OpenSMILE acoustic features and an SVM classifier yields the highest performance in speech classification, achieving an F1 score of 0.89. In contrast, the best-performing text model, using LaBSE embeddings, attained an F1 score of 0.88.

Keywords: Hate Speech Detection, Dravidian Languages, Telugu Hatespeech, Secure Cyber Physical Systems

1 Introduction

The integration of Cyber-Physical Systems (CPS) with digital social platforms requires immediate attention to develop reliable, trustworthy human-system interaction systems. The uncontrolled dissemination of hate speech content creates major threats to human safety and emotional health. The systems must implement intelligent mechanisms for detecting harmful speech to ensure security while maintaining ethical standards and creating psychologically safe digital spaces, particularly in multilingual and low-resource contexts. The spread of hate speech on social networks is a concerning problem, as it significantly impacts the psychological well-being of a person, and can lead to depressive symptoms, as highlighted in a study described in [1]. Effective detection systems must be developed because they act as a vital defense against spreading such threats. The reduction of data availability in Telugu creates significant obstacles for researchers attempting to develop sustainable hate detection frameworks. The Telugu language lacks word boundary definition as English does which generates challenges when automating detection operations because the language shows significant structural variations. Existing hate speech detection methods primarily use text-based models for analysis but the increasing use of speech-based hate content requires evaluating multi-modal detection participation in the evaluation procedure. The quantification of textual and vocal input value for hate speech detection purposes will advance more effective detection technology development.

The hate speech detection task relies on text features to extract lexical and syntactic structural and semantic meaning that identify hateful content [2]. However, text features are lacking when speakers use sarcasm, implicit hate, or context-dependent speech. Speech signals, in contrast, include prosody, tone, and acoustic variations that can identify aggression or offensive intent, which may not be verbally expressed in text alone [3].The complete dependence on speech-based models faces restrictions because of varying pronunciation patterns along with noisy backgrounds and different speaker voices. Research now needs to study the dependency between different hate speech detection modalities since their combined approaches prove capable of enhancing total performance results [4]. Modality dependency analysis serves as a useful tool to guide data collection strategies because it reveals specifically which modifications to the data will produce better results. The classification of multimodal hate speech remains important work but researchers have paid insufficient attention to how individual modalities matter before training their models together. Multimodal models are typically employed after researchers inspect individual modality results prior to application [5]. The current research assesses Telugu hate speech detection through an exploratory study of modality dependency. The research introduces a performance evaluation of multiple models and both speech and text modalities for hate detection tasks utilizing our provided dataset.

The structure of the paper is as follows: Section 2 reviews the related works in hate speech detection across different modalities and languages. Section 3 describes the dataset preparation process, including data collection, annotation, and preprocessing. Section 4 outlines the methodology involving feature extraction and model development. Section 5 and 6 presents the experimental results along with performance comparisons across different models and modalities. Section 7 discusses potential future directions for extending this research, and finally, Section 8 concludes the paper.

2 Related Works

Abreham et al. [6] addressed the challenge of scarce language resources by collecting and annotating Amharic hate speech data from YouTube. Their work produced 6,497 segments and leveraged the Google Speech-to-Text API to extract both acoustic and textual features. The expert annotation methodology applied in their study has served as a reference for developing a Telugu hate speech dataset that combines audio and text modalities. The authors of [7] explored the challenge of explainable audio-based hate speech detection by taking a step further by identifying the specific time segments in which hate speech is present. Their experiments involved investigating two approaches: An End-to-End (E2E) approach that operates on raw audio signals directly for classification and a cascade approach that involved transcribing the audio samples to text before the classification. Their findings concluded that the E2E approach outperformed the cascade approach for the accurate detection of hate content within audio samples. This emphasizes the necessity of employing audio-based techniques for the task of Hate speech detection. In another approach, the authors in [8] proposed an automated method to generate audio-based hate speech datasets via a neural text-to-speech (TTS) model. They converted 200 Toxigen text samples into 600 synthetic audio recordings that cover diverse demographics. Although their findings indicated improved performance using audio-based techniques, the artificial nature of TTS-generated data may not fully capture the natural prosody present in manually recorded speech.

Soman et al. [9] built a sentiment analysis dataset for Tamil and Malayalam by collecting 134 YouTube movie review videos. These videos were manually transcribed and annotated using a five-point sentiment scale, with strict video selection and Fleiss's Kappa for agreement verification providing important insights into annotation challenges in low-resource languages. Sumathi et al. [10] focused on text-based analysis by leveraging existing bilingual datasets in Tamil and Telugu (e.g., Tamil Binary Classification, ACTSEA, and ACTSA) for sentiment classification. Their experiments showed that a bi-directional GRU achieved the highest accuracy of 81%, followed by BiLSTM+Attention (75%), BiLSTM+CNN (66%), and IndicBERT (56%). The study highlights that models with fewer parameters can generalize well on small datasets, while more complex models like IndicBERT may struggle in bilingual contexts. Further extending the exploration of text modalities, another study [11] developed a monolingual Telugu tweet corpus annotated with hate and non-hate labels. Fine-tuned

transformer-based models such as mBERT, DistilBERT, and Indic-BERT were evaluated, with mBERT achieving an accuracy and F1 score of 98.2%. This work establishes a strong baseline for hate speech detection in resource-constrained languages.

A multimodal deep learning approach combines acoustic features with textual features to study the detection of Amharic hate speech according to the study by Abreham et al. A group of researchers extracted Word2Vec textual features together with MFCC acoustic features to conduct tests using four deep learning models that included LSTM and Bi-LSTM as well as GRU and Bi-GRU. [6].Experimental verification demonstrated that multilayered models proved more efficient than single-layered models where the Bi-LSTM implementation reached the best accuracy of 88.15% for the detection of hate speech. BILSTM proved to be the optimal architectural choice for detecting hate speech through multimodal detection due to its ability to integrate context from future and past input sequences according to the authors. The research develops [12] a three-dimensional abusive language detection method and a sentiment analysis system that employs mBERT text elements and MFCC-derived audio signals together with visual analysis through ViT. The system evaluation showed that the mBERT and MFCC models achieved the highest F1 score (0.5786) in abusive language detection, but the combined model (mBERT + ViT + MFCC) scored 0.5555. The ViT-based visual model generated the best performance in sentiment analysis with results of 0.357 for Tamil and 0.233 for Malayalam.

Prior research in hate speech detection often suffers from key limitations such as the use of synthetic speech data, a focus on non-Telugu or high-resource languages, and an overreliance on text-based approaches that miss crucial speech-related cues like tone and prosody. To overcome these issues, our study presents a manually annotated multimodal Telugu dataset collected from real YouTube content, capturing the nuances of natural speech. We systematically evaluate both individual (text and speech) using a mix of classical machine learning and transformer-based models, enabling a balanced and effective analysis tailored for low-resource settings.

3 Dataset Preparation

3.1 Data Collection

The dataset used in this study was created as part of the DravLangGuard dataset, which is a comprehensive multi-modal hate speech dataset that includes three Dravidian languages: Telugu, Tamil, and Malayalam. Our Telugu dataset includes one hour of hate audio data and one hour of non-hate audio data, along with their corresponding Telugu transcripts. The dataset is gathered from the YouTube platform, specifically from channels that have 50,000 subscribers. The procedure involved in creating this dataset is clearly shown in Figure 1. All audio samples are sampled at 44.1kHz in mono channel and stored in .wav format.

In case of hate speech (H), we focused on gathering audio samples from 4 sub classes, which are in accordance with YouTube's hate speech policy: Gender (G) based hate speech is speech against a person or a group because of their gender with a tendency for stereotypes, discrimination, or negative comments. Religion-based hate is directed towards members of a religion, creating hatred based on religion or religious

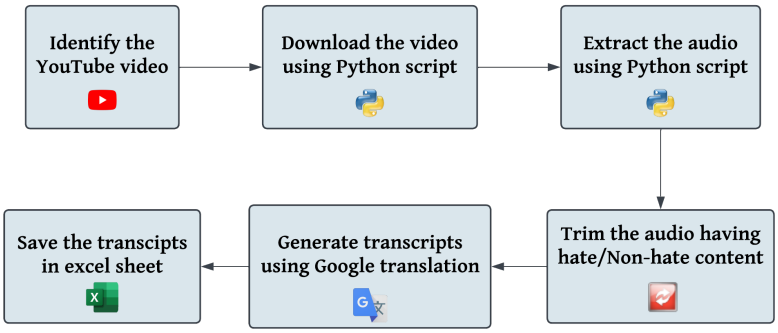

Figure 1: Overview of the data collection pipeline.

beliefs. Political/nationality (P)-based hate speech is directed at individuals or organizations based on their political or national identity. Personal Defamation (C)-based hatred is defined by personal insults and false charges intended to harm an individual's reputation. In contrast to hate speech data, we collected samples of motivating and devotional content for non-hate class. Table 1 displays the sample distribution for 5 classes and their durations.

Table 1: Distribution of data samples and their durations across each class

	Sub Class	Short Label	No. of Samples	Total duration (min)
Hate (H)	Gender	G	111	15.75
	Religion	R	82	15.49
	Political / Nationality	P	68	14.90
	Personal Defamation	C	133	14.90
Non Hate (NH)	Non Hate	N	208	60.00

3.2 Data Annotation

After collecting the audio and text samples from YouTube, we approached 3 native speakers (2 Male & 1 Female) of Telugu language who are Post Graduate degree holders and asked them to annotate the data. Firstly, they were asked to classify the data into hate and non-hate data and then asked them to categorize the hate data into 4 sub-classes as described in Section 3.1, according to YouTube's Hate Speech Policy. The average inter-annotator agreement is measured using Cohen's Kappa which is around 0.79 for our dataset.

After completing the annotation procedure, each data sample was assigned a standardized name to improve readability and the dataset structure. The naming convention

follows a uniform format and the label assigned to each sample includes the following information:

- **Speech Type:** Hate (H) / Non-Hate (NH)
- **Hate Category:** Gender (G) / Religion (R) / Political (P) / Personal Defamation (C) / Non-Hate (N)
- **Speaker Gender:** Male (M) / Female (F)
- **Parent Video Identifier:** A three-digit code representing the source video
- **Segment Identifier:** A three-digit code representing the specific segment within the parent video.

This structured naming convention ensures efficient dataset management and facilitates subsequent analysis.

4 Methodology

This paper presents a modular pipeline that was specifically engineered for each modality to handle the hate speech detection task, as demonstrated in Figure 2. The dataset was separated into training data containing 480 samples and testing data with 121 samples. The study begins with an analysis of the performance of speech modalities on the dataset to check its effectiveness in detecting hate content. For each modality, our task is to perform binary and multiclass classification.

Following the evaluation of speech modality, we then performed an ablation study using the textual data to explore their suitability for our task. We used various models and evaluated their overall performance. The following sections provide a detailed information on the models we used and their performance.

4.1 Speech feature extraction

For the speech modality, we have employed five different feature extraction methods. OpenSMILE [13], an open source toolkit, is utilized to extract robust acoustic features from speech signals. Employing the standard OpenSMILE toolkit with the ComParE 2016 configuration file, we extracted 15 distinct acoustic features, including Zero-Crossing Rate, RMS Energy, 12 MFCC coefficients, and Voicing Probability. To transform the raw features into a robust representation, we applied 10 statistical functionals to each of the 15 features. These functionals encompass measures such as the minimum (or maximum) value and its relative position, range, median, kurtosis, skewness, standard deviation, and linear regression coefficients with quadratic error, which resulted in a 150-dimensional feature vector. Using statistical features rather than the raw features, allows the model to effectively learn the non-linear dynamics of the speech data which is essential considering the limited annotated data we have.

The second feature extraction utilizes the LIBROSA library [14] to obtain complementary audio features from raw speech signals by extracting the mel spectrogram, mel frequency spectral coefficients (MFCC) and chromagrams. The time-frequency energy distribution is measured through 128 mel filter banks which generate the mel

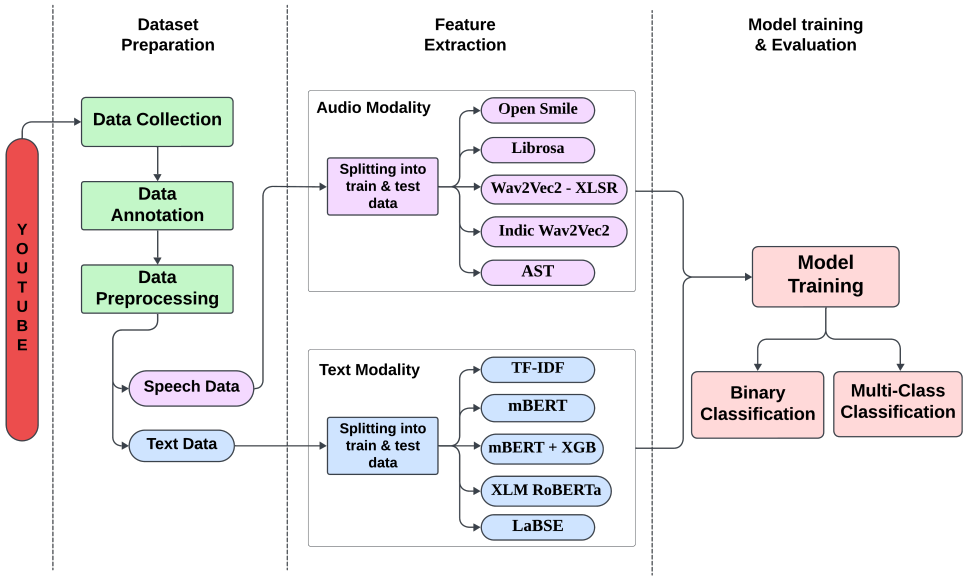

Figure 2: Overview of pipeline for Hate Speech detection.

spectrogram, while the short-term power spectrum and spectral envelope characteristics are encoded by extracting 40 MFCCs. A set of twelve chroma features exists to describe spectral energy distribution across twelve pitch classes. We obtain a suitable vector for classification by calculating the mean time axis of each feature matrix followed by horizontal concatenation of the mean vectors. The method produces an easy-to-process feature vector that summarizes global statistics of speech signals for effective and efficient classification.

A self-supervised pre-trained deep learning model called Wav2Vec2 (Cross-Lingual voice Representation) [15] is used to generate contextualized voice representations from raw waveform data. This removes the requirement for hand-crafted feature engineering, as is done in traditional approaches such as OpenSMILE or MFCCs. In this study, we employed the Wav2Vec2 xlsr-53 model, which generates 1024-dimensional feature embeddings at each time step. The resulting feature matrices measured (480, 1, 1024) for training data and (121, 1, 1024) for testing data. These embeddings were then fed into downstream classifiers, such as LSTM and CNN models, for binary and multi-class hate speech classification.

Indic-Wav2Vec2 is another self-supervised speech representation model [16] that is built on the Wav2Vec2 architecture and particularly trained on speech data from multiple Indic languages. The motivation behind using model stems from its ability to capture language-specific phonetic and prosodic features better than models trained on predominantly non-Indic datasets. We utilized the Indic Wav2Vec2-base model, which generates 1024-dimensional embeddings for each audio sample. The extracted feature matrices had dimensions (480, 1, 1024) for the training set and (121, 1, 1024) for the test set, ensuring consistency with other speech-based feature extraction methods.

The final feature extraction that we used was the Audio Spectrogram Transformer (AST), a transformer-based model designed for audio classification tasks. Unlike convolutional networks that extract local hierarchical features, AST directly processes spectrogram representations of audio samples using the attention mechanism of Vision Transformer (ViT) architecture. The model has shown strong performance in various audio-related tasks, particularly in large-scale classification problems [17]. For our study, we utilized AST base model to project 768-dimensional feature embeddings of the spectrogram representation of Telugu speech samples. The resulting feature matrices were (480, 768) in the training set and (121, 768) in the test set. Features are subsequently utilized to feed into the classification layer of the model to perform binary and multi-class classification.

5 Ablation Study: Text feature extraction

In the context of text modality, we employed five different approaches for extracting relevant features. Term Frequency-Inverse Document Frequency (TF-IDF) represents textual data through statistical calculations that measure word importance through its appearance frequency within a single document against the entire document collection. Using this method, terms are given weight based on how frequently they occur in a document but penalized for common words present in other documents to focus on more unique terms. Since our dataset is comprised of hate and non-hate speech transcriptions, we use the TF-IDF as a baseline feature extraction method and evaluate the efficacy of traditional machine learning models for the hate speech identification. We preprocessed the text data and extracted the TF-IDF features, and each transcript is represented as a high dimensional sparse vector for our study. Finally, the feature matrix was trained with the classifiers, such as Logistic Regression, Support Vector Machine (SVM) and Random Forest, for the binary and multi-class classification.

Multilingual BERT is a transformer-based language model trained across more than 100 languages, including Telugu. mBERT's ability to deal with multilingual problems is also useful for cross-lingual hate speech detection, rather than monolingual BERT models. So for this study, we fine-tuned the mBERT base model on our dataset and extracted 768 dimensional contextualized embeddings for each text input. The feature vectors produced were sent through the classification layer of the model to perform the binary and multi-class classification. Also, the extracted features are also passed to the XG Boost Classifier to predict on this set of features.

XLM RoBERTa is a RoBERTa-based multilingual model that was trained at 2.5TB of multilingual data collected from CommonCrawl. XLM-R performs better cross-lingual generalization than mBERT, and also has larger model capacity, making it a strong multilingual model for hate speech detection. Since XLM-R is pre-trained on sentence level representations, it is especially powerful in representing semantics, word order, and contextual relationships in hate speech text. We fine tuned the XLM-R on Telugu transcriptions and extracted the 1024 dimensional embedding for each input sequence and fed them into the classification layer of the model for performing the binary and multi-class classification.

The final feature extraction method we employed is LaBSE which is a cross-lingual text similarity and classification sentence embedding model. In contrast to mBERT and XLM-R, which generate token representations, LaBSE gives sentence-level embeddings that describe semantic matching across languages and therefore make it quite compelling for cross-lingual hate speech detection. We extracted 768 dimensional sentence embeddings for each text sample using LaBSE and fed them to the classification layer. Since LaBSE was designed for cross-lingual retrieval tasks, its embeddings are engineered to uphold semantic consistency across different languages, which is useful for dealing with code-mixed and transliterated text that is commonly encountered on Telugu social media.

6 Results

Our experiments with speech data evaluated various feature extraction methods and classifiers. We tested different combinations of feature extraction and classifiers for both binary and multi-class classification. Among the traditional acoustic features, OpenSmile with SVM provided a best binary classification accuracy of 91%, whereas its Gradient Boosting counterpart performed best in the multi-class case with 84% (refer to Table 2).

Table 2: Results of Speech-based Models

S.No	Model	Classifier	Binary Accuracy	Multi-class Accuracy
1	OpenSmile	SVM	91%	79%
		Gradient Boost	82%	84%
2	Wav2Vec2 - XLSR	LSTM	84%	68%
		CNN	84%	66%
3	Indic Wav2Vec2	-	90%	68%
4	AST	-	86%	81%
5	Librosa Feature Set	-	88%	61%

Deep learning-based models like Wav2Vec2-XLSR with LSTM or CNN produced uniform binary accuracies of 84%, although their multi-class accuracies were relatively lower (66–68%). Indic Wav2Vec2 worked satisfactorily in binary classification (90%) but proved to be lacking in multi-class classification (68%). The Audio Spectrogram Transformer (AST) had balanced performance with 86% binary and 81% multi-class accuracy. while LIBROSA-based features were successful in binary detection (88%) but poor at multi-class detection (61%). These findings suggest a complementary strength between traditional acoustic feature-based models and deep learning approaches—while the former excels in binary tasks, the latter tends to generalize better across multiple hate speech categories. A visual summary of these results is shown in Figure 3.

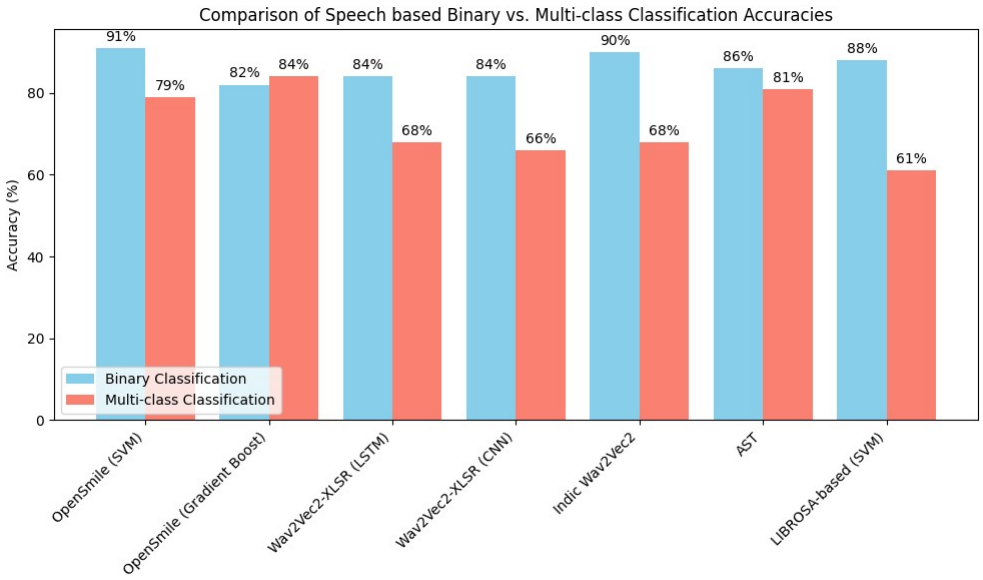

Figure 3: Overview of accuracies for Speech-based models.

While speech modalities are the main focus of this study, we also investigated text-based methods as part of our ablation study. Classical approaches with TF-IDF features and Random Forest and SVM classifiers performed poorly, having binary accuracies of 78% and multi-class accuracies below 55% (refer to Table 3).

Transformer models showed better performance. mBERT registered 83% binary and 71% multi-class accuracy, but XLM-RoBERTa outperformed them with 84% binary and 80% multi-class accuracy. LaBSE showed the best binary classification results of 89%, accompanied by a robust multi-class accuracy of 76%. Most importantly, mBERT combined with XGBoost returned decent scores (79% binary, 64% multi-class), showing that fine-tuned transformer models by themselves offer superior generalization abilities compared to standard machine learning pipelines. An overview of these text-based results is illustrated in Figure 4.

Table 3: Results of Text-based Models

S.No	Model	Classifier	Binary Accuracy	Multi-class Accuracy
1	TF-IDF	SVM	78%	55%
		Random Forest	78%	48%
2	mBERT	XG Boost	79%	64%
3	mBERT	-	83%	71%
4	XLM-RoBERTa	-	84%	80%
5	LaBSE	-	89%	76%

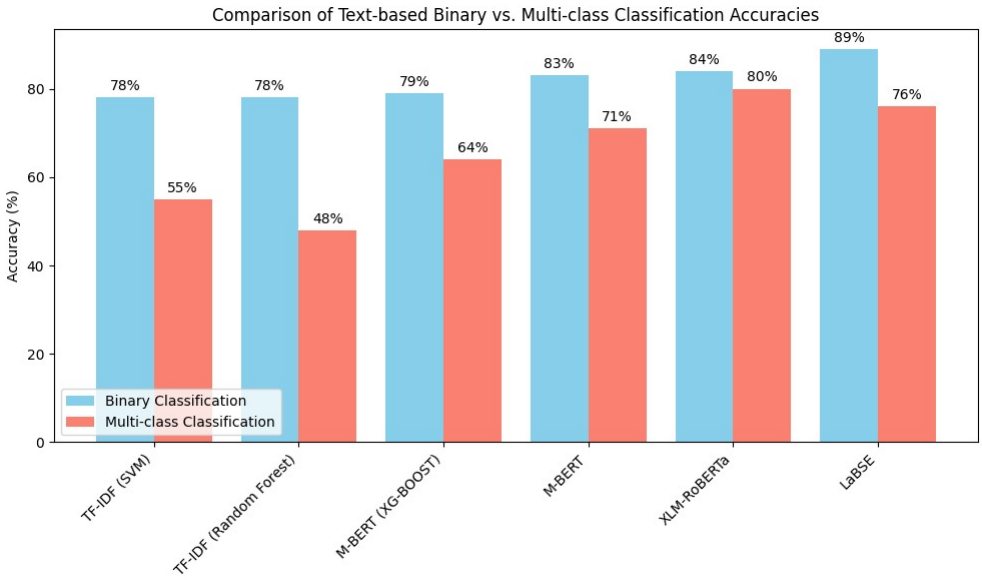

Figure 4: Overview of accuracies for Text-based models.

7 Future Prospects

This work can be extended in a number of meaningful ways. The primary focus should be on the dataset expansion, as the current version, while useful, contains only two hours of annotated data. More samples with a variety of speaker demographics and dialects in the dataset will improve the model's performance and generalizability. Future research should investigate multi-modal strategies that combine speech and textual features for analysis. The use of advanced fusion techniques such as joint embeddings and transformer-based fusion and co-attention mechanisms would help reveal more complex cross-modal relationships. The use of multi-modal methods becomes beneficial when individual modalities lack sufficient information to detect hate content in the sample.

In conclusion, our research demonstrates the potential of speech-based hate speech detection in Telugu while offering a properly curated dataset and baseline models to support future investigations in low-resource multimodal natural language processing.

8 Conclusion

This study presents a comprehensive approach to the identification of hate speech in Telugu through both speech and text modalities, with a special focus given to the speech modality based on the limited annotated Telugu speech resources. We created a new multimodal dataset of 2 hours of Telugu annotated audio-text pairs from

YouTube with a balanced hate and non-hate class distribution. Speech-based experiments examined a variety of feature extraction methods—hand-designed features and deep learning-based representations in conjunction with vanilla and neural classifiers.

Among the methods tested, OpenSMILE features with an SVM classifier performed the best for binary classification (91% accuracy and 0.89 F1-score), reaffirming the usefulness of statistical functionals for capturing speech-specific hate traits. Deep learning models such as AST and Indic Wav2Vec2 show comparable performances, especially in generalizing to multi-class conditions. These models underscore the significance of the exploitation of acoustic information in hate speech detection, particularly in a low-resource language like Telugu, particularly where textual representations can be transliterated or code-mixed.

Although text-based experiments were conducted as part of the ablation study, they revealed the strong generalization ability of sentence-level transformer models. Notably, LaBSE achieved the best binary classification accuracy (89%) among text methods, suggesting the utility of semantically rich, language-agnostic embeddings. However, speech-based detection showed slightly better robustness in binary settings, highlighting its value when textual content is ambiguous or unavailable.

References

[1] Madriaza P, B.-A.S.M.A.D.-C.L.B.E.P.D.P.S. Hassan G: Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities (2025) https://doi.org/10.1002/cl2.70018.PMID:39822240;PMCID:PMC11736891

[2] Jahan, M.S., Oussalah, M.: A systematic review of hate speech automatic detection using natural language processing (2023) https://doi.org/10.1016/j.neucom.2023.126232

[3] Tembe, L.A., Anand Kumar, M.: Hate speech detection using audio in portuguese language. In: Speech and Language Technologies for Low-Resource Languages (2024)

[4] Jyothish Lal, G., Premjith, B., Chakravarthi, B.R., Rajiakodi, S., Bharathi, B., Natarajan, R., Rajalakshmi, R., *et al.*: 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, pp. 114–122 (2025)

[5] Mohan, J., Mekapati, S.R., B, P., G, J.L., Chakravarthi, B.R.: 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 (2025)

[6] Debele, A.G., Woldeyohannis, M.M.: Multimodal amharic hate speech detection using deep learning. In: 2022 International Conference on Information and

Communication Technology for Development for Africa (ICT4DA), pp. 102–107 (2022). https://doi.org/10.1109/ICT4DA56482.2022.9971436

[7] An, J., Lee, W., Jeon, Y., Ok, J., Kim, Y., Lee, G.G.: An investigation into explainable audio hate speech detection. In: Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 533–543. Association for Computational Linguistics, ??? (2024). https://doi.org/10.18653/v1/2024.sigdial-1.45 . https://aclanthology.org/2024.sigdial-1.45/

[8] Bhesra, K., Shukla, S.A., Agarwal, A.: Audio vs. text: Identify a powerful modality for effective hate speech detection. In: The Second Tiny Papers Track at ICLR 2024 (2024). https://openreview.net/forum?id=dD2e3aCEcO

[9] Chakravarthi, B.R., K, J.P.P., B, P., Soman, K.P., Ponnusamy, R., Kumaresan, P.K., Thamburaj, K.P., McCrae, J.P.: DravidianMultiModality: A Dataset for Multi-modal Sentiment Analysis in Tamil and Malayalam (2021). https://arxiv.org/abs/2106.04853

[10] Sumathi, D., Gowtham, B., Naveen, K., Subramani, H.: Sentiment classification on tamil and telugu text using rnns and transformers. (2021). https://doi.org/10.1109/ICTAI53825.2021.9673365

[11] Khanduja, N., Kumar, N., Chauhan, A.: Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. Systems and Soft Computing **6**, 200112 (2024) https://doi.org/10.1016/j.sasc.2024.200112

[12] Barman, S., Das, M.: hate-alert@DravidianLangTech: Multimodal abusive language detection and sentiment analysis in Dravidian languages. In: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, pp. 217–224 (2023). https://aclanthology.org/2023.dravidianlangtech-1.31/

[13] Eyben, F., Weninger, F., Groß, F., Schuller, B.: openSMILE: The Munich Open-Source Large-Scale Multimedia Feature Extractor (Version 3.0.0). Accessed: 2024-05-23 (2020). https://github.com/audeering/opensmile/releases/tag/v3.0.0

[14] McFee, Raffel, Liang, P.W. Ellis, McVicar, Battenberg, Nieto: librosa: Audio and Music Signal Analysis in Python. In: Huff, Bergstra (eds.) Proceedings of the 14th Python in Science Conference, pp. 18–24 (2015). https://doi.org/10.25080/Majora-7b98e3ed-003

[15] Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised Cross-lingual Representation Learning for Speech Recognition (2020). https://arxiv.org/abs/2006.13979

[16] Chadha, H.S., Gupta, A., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R.,

Raghavan, V.: Vakyansh: ASR Toolkit for Low Resource Indic languages (2022). https://arxiv.org/abs/2203.16512

[17] Gong, Y., Chung, Y.-A., Glass, J.: AST: Audio Spectrogram Transformer (2021). https://arxiv.org/abs/2104.01778