Towards Safer Online Spaces: Deep Learning for Hate Speech Detection in Code-Mixed Social Media Conversations

Supriya Chanda*
supriyachanda.rs.cse18@itbhu.ac.in
Indian Institute of Technology (BHU) Varanasi
Varanasi, Uttar Pradesh, INDIA

Abhishek Dhaka*
abhishek.dhaka7340@gmail.com
B. K. Birla Institute of Engineering & Technology
Pilani, Rajasthan, INDIA

Sukomal Pal
spal.cse@itbhu.ac.in
Indian Institute of Technology (BHU) Varanasi
Varanasi, Uttar Pradesh, INDIA

ABSTRACT

In the midst of the widespread adoption of technology, particularly among younger generations, the increasing prevalence of hate speech online has become a pressing global concern. This research paper aims to address this urgent issue by conducting a thorough investigation into hate speech detection in Hindi-English code-mixed data. Existing research has largely approached hate speech recognition as a text classification problem, focusing on predicting the class of a message based solely on its textual content. Our task, however, delves into the classification of hateful content disseminated through tweets, comments, and replies on Twitter, taking into account the contextual intricacies inherent in social media communication. In this context, contextual nuances play a crucial role in understanding communication dynamics. By employing state-of-the-art deep learning techniques tailored to the unique linguistic characteristics of each language, this research makes a significant contribution to the development of robust and culturally sensitive hate speech detection systems. Such systems are essential for creating safer online environments and promoting cross-cultural understanding.

Warning: The content of this paper may contain offensive material, reader discretion is advised.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

KEYWORDS

Hate Speech Identification, Code-Mixing, Hindi-English, mBERT, Sentence BERT

ACM Reference Format:
Supriya Chanda, Abhishek Dhaka, and Sukomal Pal. 2024. Towards Safer Online Spaces: Deep Learning for Hate Speech Detection in Code-Mixed Social Media Conversations. In *16th ACM Web Science Conference (Websci Companion '24), May 21–24, 2024, Stuttgart, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3630744.3663610

*Both authors contributed equally to this research.

Websci Companion '24, May 21–24, 2024, Stuttgart, Germany

ACM ISBN 979-8-4007-0453-6/24/05
https://doi.org/10.1145/3630744.3663610

1 INTRODUCTION

The advent of social media has brought together individuals from diverse racial and educational backgrounds, facilitated by the widespread accessibility of the Internet. However, alongside the positive aspects, there has been a proliferation of offensive language and hate speech within online communities. The content generated by users often serves as a fertile ground for the dissemination of such harmful rhetoric, resulting in a decline in user morale and potential psychological repercussions such as trauma, anxiety, and post-traumatic stress disorder (PTSD). Addressing this issue is paramount to fostering civil discourse, promoting adherence to community guidelines, and ensuring online safety and inclusivity.

Code-mixing, the phenomenon wherein individuals use multiple languages within the same conversation, is particularly prevalent in informal online settings such as social media. In linguistically diverse nations like India, individuals frequently employ regional languages alongside English for communication on digital platforms. A study by Grover et al. [10] examined patterns of English-Hindi code-mixing and slang usage on social networks, revealing a strong preference for specific languages among bilingual individuals and highlighting the influence of profanity on code-switching behaviors.

While the identification of hate speech has been extensively studied in monolingual contexts, processing and analyzing code-mixed data present unique challenges due to the incorporation of multiple languages and linguistic variations. Academics and practitioners must develop effective tools and methodologies to handle such data efficiently, particularly in the context of hate speech identification, given its prevalence on social media platforms.

In response to these challenges, there has been a growing interest in the development of automated systems for identifying hate content in code-mixed social media data. Such tools have the potential to significantly contribute to the fight against hate speech and foster a more inclusive and respectful online community.

This study focuses on identifying hate content in code-mixed social media data, treating it as a text classification problem and exploring various deep learning methodologies for resolution. The datasets utilized are sourced from the FIRE 2022 and FIRE 2023 (Forum for Information Retrieval Evaluation) Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC), comprising Hindi-English code-mixed data.

The primary contribution of the paper lies in addressing the challenge of identifying hate speech in Hindi-English social media

conversations within the framework of the HASOC shared task on Identifying Conversational Hate-Speech in Code-Mixed Languages. We leverage deep learning models like mBERT to analyze the context of posts, comments, and replies, addressing the unique challenges of hate speech detection in code-mixed social media conversations. Our approach emphasizes contextual understanding and semantic analysis, enabling more effective identification of both standalone and contextual hate speech and offensive content. Additionally, we implement rule-based classification to combine model outputs. Through our novel methodologies and ensemble model strategy, we aim to advance the field of hate speech detection and promote the creation of safer online spaces.

The task descriptions are outlined below.

1.1 Tasks Description

The objective of the Hate Speech and Offensive Content Identification (HASOC) initiative was to provide a standardized platform for the automated detection of hate speech and objectionable content across various languages within social media posts. Over the past four years, HASOC has organized numerous tasks across different linguistic contexts. Notably, in 2023, a task focusing on conversational code-mixed data for hate speech identification was introduced.

The delineation of tasks within this study is structured as follows:

Task 2: This task known as the Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL), addresses the challenge of identifying hate speech and offensive content in code-mixed conversations on social media. The task is divided into two subtasks.

Task 2a: ICHCL HINGLISH Codemix Binary Classification.

Task 2a, part of the Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) challenge, focuses on the detection of hate speech and offensive content within code-mixed conversations occurring on social media platforms. Code-mixed text refers to conversations where multiple languages are interwoven within a single dialogue. This task involves binary classification. Participants are tasked with determining whether a tweet, comment, or reply contains hate speech, offensive language, or profanity (HOF), or if it is devoid of such content and considered non-hate and non-offensive (NOT). Notably, this classification should account for both the content of the individual tweet or comment and any support for hate expressed within the parent tweet.

- **(NOT) Non Hate-Offensive** - This post does not contain any hate speech, profanity, or offensive content.
- **(HOF) Hate and Offensive** - Contains Hate, offensive, and profane words.

Task 2b: ICHCL HINGLISH Codemix Multiclass Classification.

Task 2b, as part of the ICHCL challenge, extends the identification of hate speech in code-mixed conversations by delving into specific forms of hate expression. In this task, participants are entrusted with the classification of conversational tweets, comments, and replies that possess hierarchical-structured data. They are required to identify whether the content falls into distinct categories: Standalone hate (SHOF), Contextual hate (CHOF), which pertains to content supporting hate expressed in the parent conversation, or non-hate (NONE).

- **(SHOF) Standalone Hate** - Contains Hate, offensive, and profane words in itself.
- **(CHOF) Contextual Hate** - Comment or reply of a tweet supports the hate, offense, and profanity expressed in its parent tweet. This includes expressing apparent hatred and endorsing the hatred with positive sentiment.
- **(NONE) Non-Hate** - Any form of Hate speech, profane, offensive content is not present.

Table 1: Example tweets from the HASOC2023 dataset for all classes

Language	Sample tweet from the class	Task 2a	Task 2b
Hinglish	@Joydas @NSaina 2 rs k liye tweet kiya isne samjho bhai national hero hogi phle ab to izzat gawa di.. Andhbhakt ban gayi didi	HOF	CHOF
Hinglish	The only thing I want to say to the Islamophobic State.#SharjeelImam #Shaheen_Bagh #JNU #Chakaa_jaam_is_not_sedition #releaseallpoliticalprisoners @URL	HOF	SHOF
Hinglish	@AUTHOR CAA also not against muslims	NOT	NONE

The remaining sections of the paper are structured as follows. In Section 2, a brief outline of some previous attempts is provided. The dataset description is presented in Section 3. Different model framework are discussed in Section 4. Results and discussions are presented in Section 5, and the conclusion is provided in Section 6.

2 RELATED WORK

The identification of hate speech and offensive content has attracted considerable attention within both academic and commercial spheres. Although a substantial volume of research has predominantly focused on English, owing to its widespread global usage, there exists an urgent requirement for equivalent corpora in other languages to comprehensively tackle this issue. Several studies have delved into the varied aspects of offensive content, such as *abusive language* [16, 18], *cyber-aggression* [11], *cyber-bullying* [5, 26], and *toxic comments or hate speech* [7, 8, 12]. A brief overview of some notable works in these areas is provided.

- *Hate Speech Identification :* Hate speech, a pervasive challenge, has been systematically categorized into various types based on the nature of its textual content. Diverse datasets have been curated to cater to these distinct categories of hate speech. Notably, a common dataset [14] has served as a foundation for identifying hate speech and profanity, with recent work by Davidson et al. [7] making use of a dataset comprising nearly 24,000 labeled tweets.
- *Offensive Content and Cyberbullying :* The broader domain of offensive content includes abusive language [11], cyber-aggression, cyber-bullying, and toxic comments. Previous investigations have employed techniques such as sentiment analysis, topic modeling [26], and user-related features [5] to tackle this multifaceted problem.

Efforts have extended beyond English, with endeavors in languages including German [15, 21], Spanish, Arabic [16, 17], Greek [19], Slovene [9] and Chinese [24]. Mubarak et al. [16] introduced a collection of profane terms, known as SeedWords (SW), and applied the Log Odds Ratio (LOR) to individual word unigrams and bigrams. Saroj et al. [23] adopted a Support Vector Machine (SVM) approach

alongside TF-IDF features, targeting hate speech and offensive language in Arabic and Greek.

In recent years, initiatives like HASOC [15] and GermEval [25] have spotlighted the importance of addressing hate speech detection in various languages and contexts. Dravidian LangTech [2], for example, focused on detecting offensive language in a code-mixed dataset comprising Tamil–English, Malayalam–English, and Kannada–English. The application of multilingual models, including BERT variants and IndicBERT, has shown promise in this regard. Transfer learning has shown potential in enhancing offensive language recognition, particularly in code-mixed contexts. Researchers have leveraged transfer learning from English datasets to improve offensive language recognition in code-mixed Kannada [22], Malayalam [20], and Tamil [27].

Detecting hate speech within conversational Hindi-English code-mixed data introduces additional complexities owing to the conversational nature inherent in such content. The hierarchical structure comprising posts, comments, and replies mandates a nuanced approach, necessitating a spectrum of techniques ranging from unified text treatment to innovative hierarchical neural network architectures. Each post may engender multiple comments, and each comment may elicit several replies. Within the domain of English-Hindi data, every component of the tuple can exhibit code-mixing between Hindi and English, solely English expression, exclusively Hindi communication, romanized Hindi representation, or a fusion thereof. Consequently, intricate input patterns manifest. The assignment of labels to replies or comments is notably influenced by the contextual information imparted by the parent text. To address this, Chanda et al. [4] treated all the post, comments, and replies as a single unified text and applied a pre-trained multilingual BERT model. To maintain the context of post to comments and reply, Chanda et al., [3] concatenate. Bagora et al., [1] proposed a novel hierarchical neural network architecture, while Madhu et al., [13] employed a pipeline consisting of an LSTM classifier followed by a fine-tuned SentBERT model.

3 DATASET

In this study, we employed the datasets from HASOC 2023, graciously provided by the organizers of FIRE 2023. The organizers supplied the training data for the task. For the final evaluation, they made the test data available, requiring participants to submit prediction files for each data sample.

Task 2's dataset departed from the aforementioned format, comprising conversational threads potentially containing instances of hate speech and offensive content. Notably, identifying such content was not straightforward from individual comments or replies; Instead, identifying such content necessitated an understanding of the contextual parent content.

The corpus collection and class distribution is shown in Table 2.

3.1 Preprocessing

Social media data is notably informal and prone to noise due to the colloquial style prevalent in Twitter conversations. This inherent attribute presents a potential challenge to the precision of processing techniques. As a result, it has been considered essential to subject all data to preprocessing procedures aimed at alleviating the influence of less informative textual elements.

Table 2: Statistical overview of the Training Data and Test Data

Task-2A					
Data	Language	# Sentences	NONE	HOF	
Train	Hindi-English	12998	6425	6573	
Test	Hindi-English	998	-	-	
Task-2B					
Data	Language	# Sentences	NONE	SHOF	CHOF
Train	Hindi-English	5910	2873	1986	1051
Test	Hindi-English	998	-	-	-

In this task, the expansion of emojis and hashtags was conducted as part of the preprocessing pipeline. Hereafter, we offer a detailed enumeration of the preprocessing steps that were implemented.

- Perform cleaning by removing usernames, punctuation and URLs, mentions and hashtags.
- Use ekphrasis which is a text processing tool, geared towards text from social networks, such as Twitter or Facebook. ekphrasis performs normalizing hashtags (for example, "#BlackLivesMatters" is segmented into "Black", "Lives",and "Matters").

For the binary classification task, the 'HOF' labels have been converted to integer '1', representing instances of harmful or offensive content, while the 'NOT' labels have been converted to integer '0', indicating non-harmful content.

4 METHODOLOGY

This section outlines the methodology employed for Task 2a and Task 2b of the HASOC shared task, focusing on the Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL). Task 2a presents a distinct challenge, necessitating the binary classification of conversational tweets with tree-structured data to ascertain whether the content contains hate speech, offensive language, or profanity (HOF), or if it falls under the category of non-hate and offensive (NOT). Unlike conventional classification problems, we abstained from reliance on pre-trained models due to the unique nature of this task, wherein the contextual relevance of preceding posts or replies assumes paramount importance. Each classification decision was contextually driven, with comments evaluated in the context of their parent posts, and replies contextualized within the broader framework of the main post, alongside the specific comment to which the reply pertains.

4.0.1 Fine-tuning mBERT. This section provides a detailed account of the fine-tuning methodology employed for Task 2a. Given the specialized nature of this task, a fine-tuning approach was chosen over the utilization of pre-trained models, which is customary for general classification problems.

In the first phase of fine-tuning mBERT we constructed a specialized dataset comprising anchor points and their corresponding positive or negative pair. If two tweets have same true labels then

the pair was labeled as 0 otherwise pair was labeled 1 (see Figure 1). For this phase of fine tuning mBERT, we employed Siamese network.

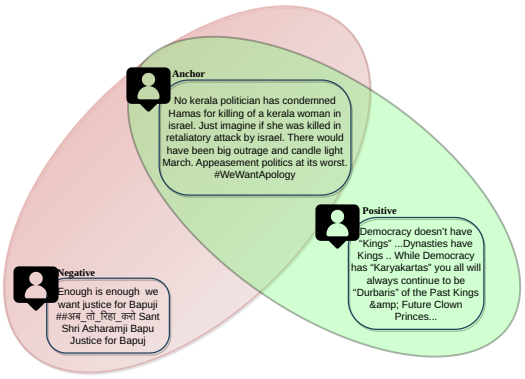

Figure 1: Creation of datasets using anchor-positive and anchor-negative pair

The fine-tuning process entailed initializing mBERT with pre-existing weights and subsequently refining it on our specialized dataset created for Siamese network. Multiple training iterations were executed, leveraging the output from mBERT, which comprised 768-dimensional contextual embeddings, as input to a dense layer featuring 128 units.

To train the fine-tuned model effectively, the Contrastive loss was used. The Contrastive loss function is defined in Equation 1. Due to computational constraints, a batch size of 8 was chosen, and experimentation revealed that a learning rate of 0.001 was optimal for fine-tuning the model.

$$\mathcal{L}(x_1, x_2, y) = \frac{1}{2}(1-y) \cdot d(x_1, x_2)^2 + \frac{1}{2} y \cdot \max(0, m - d(x_1, x_2))^2 \tag{1}$$

The decision to utilize a Siamese network was aimed at enhancing the model's ability to discern similarities between hate speech and non-hate speech. This choice was motivated by the presence of both languages in Roman script within our dataset. There are instances where hate speech does not explicitly target individuals; for example, if the post, comment, or reply contains sarcasm or expresses support for hate without explicitly using any offensive words. Therefore, leveraging the Siamese network aids in the learning process.

4.0.2 Submission for Task 2a : Expanding upon the fine-tuned mBERT model, we augmented our primary architecture to enhance the classification of hate speech and offensive content in Task 2a of the HASOC shared task. This architecture synergistically leveraged the capabilities of mBERT with the sentence transformer [13], with a maximum token length of 160 tokens taken into consideration.

In our model architecture, we adhered to the structure of mBERT, integrating a sequence of LSTM layers. Specifically, we employed a bidirectional LSTM with 512 units, succeeded by two unidirectional LSTM layers with 512 and 256 units, respectively. The output from the final LSTM layer and the output from SentenceBERT were concatenated. Because dataset encompassed three distinct data types: Posts, comments, and replies, each characterized by its own unique attributes. To accommodate these variations, our architecture seamlessly integrated all three data channels, as shown in the architectural diagram. Post, comment and reply were multiplied with their important score (α for post, β for comment and γ for reply). Afterwords a feedforward network was introduced comprising a dense layer with 1024 neurons, followed by a dropout layer with a probability of 0.3, and subsequent dense layers with 256 and 32 neurons. Finally, a single-layer neuron with a sigmoid activation function was utilized for the output. This configuration was devised to optimize both contextual and semantic understanding within our research framework.

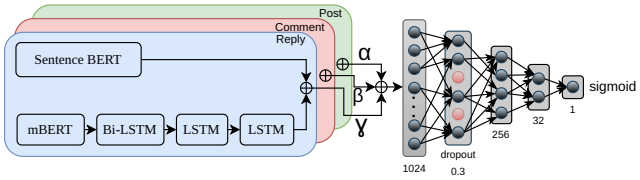

Figure 2: Architecture of proposed methodology for Task 2A

For training purposes, the data was divided into an 80% training set and a 20% validation set. A batch size of 8 was selected during training. To make classification decisions, a threshold of 0.5 was applied to the model's output probabilities.

Throughout the training phase, 50 epochs were executed with a learning rate of 0.09, utilizing the Stochastic Gradient Descent (SGD) optimizer. We got a validation accuracy of 83.19%, underscoring the efficacy of our architecture in identifying hate speech and offensive content within code-mixed conversations on social media.

4.0.3 Submission for Task 2b : An ensemble approach was employed, leveraging the collaborative operation of two distinct models. The first model utilized `hindi-codemixed-abusive-MuRIL` [6], specifically selected for its proficiency in identifying abusive language in code-mixed data within sentences, as evidenced by its applicability to the HateSpeech Hindi Code mixed Abusive MuRIL dataset. For our task which might contain hate speech without an explicit use of abusive word could still be a HOF because of presence of sarcastic comments to a race, people or gender etc. For this we need an additional classification model (see Figure 3) to better understanding the context of post, comment and reply.

For this purpose, we used a second model which underwent a comprehensive architectural transformation, distinguishing itself from the initial model. This architectural evolution encompassed the integration of several pivotal elements. It commenced with the inclusion of word-level embeddings from MuRIL, subsequently passed to a Bidirectional Long Short-Term Memory (BiLSTM) layer featuring 512 units, thereby enabling the model to effectively capture bidirectional contextual information. After bi-LSTM two LSTM

each with 256 units was unit to capture sequential patterns. Expanding upon the architectural framework, a dense layer comprising 32 units was incorporated, facilitating the extraction of features. This was succeeded by the inclusion of a final dense layer housing a single neuron and employing the sigmoid activation function, primarily responsible for generating the second model's output. Throughout the training phase, the model was fine-tuned with a learning rate of 5e-6 using a batch size of 8. It should be noted that throughout the training process the last two layers of MuRIL used in second model were also fine tuned.

The ensemble strategy relied on the synergistic collaboration between these two models, each contributing its unique strengths to effectively address the classification task. The output from both the models were added. If total sum is >=1 then the output of ensemble model is 1 (means abusive) else 0 (means not abusive) because both model give output from 0 to 1 so if a sentence have either just abusive word (in hindi or English written in any script) or if it just contain a hate speech without any abusive word in both cases total sum is greater than 1 making the final output of ensemble model as HOF.

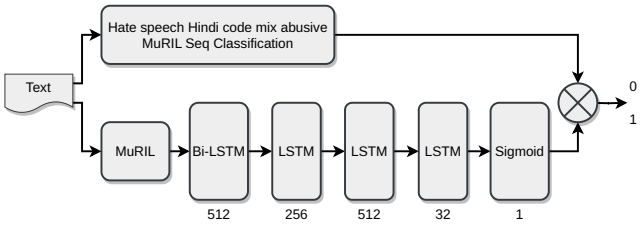

Figure 3: Architecture of proposed methodology for Task 2B

Now for the final prediction, a rule-based approach was adopted between the ensemble model given here and the model stated in 2a.The Table 3 describes the rule-based approach used to get final prediction.

Table 3: Rule Based Approach for task 2b-Method 2

Task 2a Model	Task 2b Model	Final Prediction
NOT	NOT	NOT
NOT	HOF	SHOF
HOF	NOT	CHOF
HOF	HOF	CHOF

The process involved evaluating predictions from both models to ascertain the definitive classification. For instance, in cases where model 2a predicted "NOT" while model 2b predicted "HOF," the final prediction was categorized as "SHOF." Conversely, if model 2a predicted "HOF," regardless of model 2b's prediction, the output was designated as "CHOF." This underscores the precedence of contextual hate over standalone hate within our classification framework.

5 RESULTS AND DISCUSSION

The model was validated on the training and development sets due to the limited amount of data available for training. Subsequently, the prediction file was submitted on the test data to obtain the final results.

5.1 Results

Task 2 focused on Hindi-English code-mixing, and our team's performance was notable. Table 4 and 5 shows our official performances on the test data as shared by the organizers. In subtask 2A, we achieved a Macro F_1 score of 0.70. In subtask 2B, we achieved a score of 0.56. The factor contributing to reduced score for task 2B, is the limited training data for this task as compared to task 2A. "IRLab@IITBHU" demonstrated competitive performance in both subtasks.

Table 4: ICHCL Task 2A results

Rank	Team Name	F_1 score	Recall	Precision
1	FiRC-NLP	0.8079	0.8074	0.8084
2	**IRLab@IITBHU**	**0.7008**	**0.6995**	**0.7026**
3	Chetona	0.6155	0.6143	0.6253
4	AiAlchemists	0.6147	0.6082	0.6335
5	MUCS_3	0.4347	0.500	0.3846
6	HASOC	0.3743	0.500	0.2991

Table 5: ICHCL Task 2B results

Rank	Team Name	F_1 score	Recall	Precision
1	FiRC-NLP	0.6541	0.6717	0.6433
2	**IRLab@IITBHU**	**0.5631**	**0.5668**	**0.5687**
3	AiAlchemists	0.3824	0.3921	0.3920
4	HASOC	0.2495	0.3333	0.1994
5	Chetona	0.1726	0.1588	0.2079

5.2 Discussion

In the context of this Task, which involved the challenging domain of Hindi-English code-mixing, the team embarked on a series of experiments to optimize the approach. One crucial aspect explored was the weighting of different components—namely, the post, comments, and replies—when making predictions. It was recognized that assigning the appropriate weightage to these components could greatly enhance the model's understanding of context. After experimenting with all values between 0.1 to 1.0 with step size of 0.1, it was discovered that the optimal combination of importance score was achieved when assigning α (contextual weighting coefficients for post) a value of 0.3, β (contextual weighting coefficients for comment) a value of 0.1, and γ (contextual weighting coefficients for reply) a value of 0.3. Top three experimented values are mentioned below in table.

This finding essentially implied that when predicting on reply data, it was imperative to afford equal importance to both the post and reply elements, while assigning comparatively less weight to comments. This strategic allocation of weights allowed for the effective capture and leverage of contextual nuances within the

Table 6: Top 3 Experimented Values of α, β, and γ

α	β	γ	Val Accuracy
0.3	0.1	0.3	83.489
0.3	0.1	0.7	83.114
0.3	0.1	0.1	82.926

data, ultimately resulting in the optimal scores achieved in this Task.

We experimented different number of LSTMs ranging from 1 to 4. It was observed that there was very slight improvement with increasing number of LSTM vs computation cost so using 1 biLSTM and 2 LSTM seems to be the best choice when constrained with computation resources.

In our quest for the optimal optimizer, our experimentation indicated that Stochastic Gradient Descent (SGD) with a slightly higher learning rate converges more rapidly. Conversely, the AdamW optimizer with a higher learning rate exhibited a zig-zag convergence pattern. Notably, AdamW performed optimally with a lower learning rate, typically around 1e-5. However, it is important to recognize that SGD with a marginally higher learning rate can be a pragmatic choice for quick model testing, particularly in the context of Transformer-based models. This approach provides insights into a model's convergence tendencies before committing to more computationally intensive optimization methods.

6 CONCLUSION

Throughout our investigation into the detection of conversational hate speech, several insights have come to light. The challenge inherent in this task, namely the management of text inputs of variable lengths, prompted the adoption of a strategic approach: distinct feature extraction from posts, comments, and replies. This pragmatic methodology effectively addressed the linguistic diversity within our dataset, thereby mitigating potential limitations associated with conventional models such as BERT, which impose a maximum token limit of 512 tokens for both input and output.

ACKNOWLEDGMENTS

We extend our sincere appreciation to the organizers of the HASOC competition for orchestrating this engaging shared task and datasets. Our investigation profited from meticulously curated data sets with minimal spelling discrepancies, thereby facilitating a more streamlined approach to the classification task.

REFERENCES

[1] Aditi Bagora, Kamal Shrestha, Kaushal Maurya, and Maunendra Sankar Desarkar. 2022. Hostility Detection in Online Hindi-English Code-Mixed Conversations. In *14th ACM Web Science Conference 2022*. 390–400.

[2] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv, 133–145. https://aclanthology.org/2021.dravidianlangtech-1.17

[3] Supriya Chanda, Sacchit Sheth, and Sukomal Pal. 2022. Coarse and Fine-Grained Conversational Hate Speech and Offensive Content Identification in Code-Mixed Languages using Fine-Tuned Multilingual Embedding.

[4] Supriya Chanda, S Ujjwal, Shayak Das, and Sukomal Pal. 2021. Fine-tuning Pre-Trained Transformer based model for Hate Speech and Offensive Content Identification in English, Indo-Aryan and Code-Mixed (English-Hindi) languages.

[5] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Advances in Information Retrieval*, Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 693–696.

[6] Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data Bootstrapping Approaches to Improve Low Resource Abusive Language Detection for Indic Languages. *arXiv preprint arXiv:2204.12543* (2022).

[7] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *CoRR* abs/1703.04009 (2017). arXiv:1703.04009 http://arxiv.org/abs/1703.04009

[8] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15 Companion)*. Association for Computing Machinery, New York, NY, USA, 29–30. https://doi.org/10.1145/2740908.2742760

[9] Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 46–51. https://doi.org/10.18653/v1/W17-3007

[10] Jeenu Grover, Prabhat Agarwal, Ashish Sharma, Mayank Sikka, Koustav Rudra, and Monojit Choudhury. 2017. I may talk in English but gaali toh Hindi mein hi denge: A study of English-Hindi Code-Switching and Swearing Pattern on Social Networks. IEEE.

[11] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1–11. https://www.aclweb.org/anthology/W18-4401

[12] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (Bellevue, Washington) *(AAAI'13)*. AAAI Press, 1621–1622.

[13] Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications* 215 (2023), 119342. https://doi.org/10.1016/j.eswa.2022.119342

[14] Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. *CoRR* abs/1712.06427 (2017). arXiv:1712.06427 http://arxiv.org/abs/1712.06427

[15] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya 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 Annual Meeting of the Forum for Information Retrieval Evaluation* (Kolkata, India) *(FIRE '19)*. Association for Computing Machinery, New York, NY, USA, 14–17. https://doi.org/10.1145/3368567.3368584

[16] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 52–56. https://doi.org/10.18653/v1/W17-3008

[17] Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic Offensive Language on Twitter: Analysis and Experiments. *arXiv preprint arXiv:2004.02192* (2020).

[18] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 145–153. https://doi.org/10.1145/2872427.2883062

[19] Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

[20] Tharindu Ranasinghe, Sarthak Gupte, Marcos Zampieri, and Ifeoma Nwogu. 2020. WLV-RIT at HASOC-Dravidian-CodeMix-FIRE2020: Offensive Language Identification in Code-switched YouTube Comments. *CoRR* abs/2011.00559 (2020). arXiv:2011.00559 https://arxiv.org/abs/2011.00559

[21] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *CoRR* abs/1701.08118 (2017). arXiv:1701.08118 http://arxiv.org/abs/1701.08118

[22] Siva Sai and Yashvardhan Sharma. 2021. Towards Offensive Language Identification for Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv, 18–27. https://aclanthology.org/2021.dravidianlangtech-1.3

[23] Anita Saroj, Supriya Chanda, and Sukomal Pal. 2020. IRlab@IITV at SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media Using SVM. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, Barcelona (online), 2012–2016. https://aclanthology.org/2020.semeval-1.265

[24] Hui-Po Su, Zhen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 18–24. https://doi.org/10.18653/v1/W17-3003

[25] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. Austrian Academy of Sciences, Vienna, Austria, 1 – 10. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-84935

[26] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Montreal, Canada) *(NAACL HLT '12)*. Association for Computational Linguistics, USA, 656–666.

[27] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1415–1420. https://doi.org/10.18653/v1/N19-1144