International Journal of Engineering in Computer Science

**E-ISSN:** 2663-3590
**P-ISSN:** 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2025; 7(1): 07-14
Received: 01-11-2024
Accepted: 06-12-2024

**Manish Joshi**
Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

**Dhirendra Pandey**
Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

**Vandana Pandey**
Department of Computer Application, GIHSM, Lucknow, Uttar Pradesh, India

**Mohd Waris Khan**
Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India

**Corresponding Author:**
**Manish Joshi**
Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

# A fusion framework for Hinglish cyberbullying detection using mBERT and FastText

**Manish Joshi, Dhirendra Pandey, Vandana Pandey and Mohd Waris Khan**

**DOI:** https://doi.org/10.33545/26633582.2025.v7.i1a.149

## Abstract

Cyberbullying in Hinglish, a linguistic fusion of Hindi and English widely utilised on social media, poses considerable issues due to its distinct linguistic features. a number of current detection systems are unable to adequately represent the complexity of Hinglish text, which produces less than ideal outcomes. This paper presents the Hinglish Fusion Framework for Cyberbullying Detection, a method addressing this problem by using advanced natural language processing techniques. The method lets the system detect both semantic and syntactic peculiarities of Hinglish text by combining the contextual strength of a fine-tuned BERT model with the efficiency of FastText embeddings. The framework uses a dual-stream design whereby FastText concentrates on subword-level linguistic information and BERT processes contextual embeddings. Classification performance is improved by a weighted ensemble of outputs derived from these models. Evaluated on a Hinglish cyberbullying dataset, the framework showed notable gains in precision, recall, and F1-score when compared to traditional models. With a scalable and strong solution, more inclusive and efficient moderation tools in multilingual and code-mixed environments are made possible. This paper emphasises the need of hybrid strategies for addressing the difficulties of cyberbullying detection in linguistically varied fields with limited resources.

**Keywords:** Cyberbullying, deeplearning, machinelearning, social media, fusion

## Introduction

The rapid spread of social media networks has resulted in a boom in online interaction, hence strengthening ties with information exchange on hitherto unprecedented proportions. But this connectedness has also led to cyberbullying, a ubiquitous problem seriously affecting people's mental health [1]. In multilingual areas like India, where Hinglish a mix of Hindi and English has become a major means of online expression detecting and combating cyberbullying is especially difficult. With its code-switching, casual tone, and frequent use of slang, Hinglish presents particular linguistic difficulties that render conventional natural language processing (NLP) techniques insufficient for identifying negative content.

Currently in use cyberbullying detection technologies can rely on models built on monolingual datasets, mostly in English, which misses the subtleties of Hinglish text. Low accuracy and high misclassification rates follow from these models' difficulties with linguistic diversity, context, and semantic richness. Moreover, Hinglish lacks extensive annotated datasets, which aggravates the challenges in building strong machine learning models. We present a fresh framework using advanced embeddings and attention techniques to efficiently identify cyberbullying in Hinglish text in order to handle these difficulties.

This framework's foundation is its mix of two complimentary representation approaches: FastText's lexical embeddings and multilingual BERT (mBERT) contextualised embeddings. By using subword tokenisation, transformer-based model mBERT performs well in capturing the contextual semantics of code-switched text. Conversely, FastText a word embedding model based on character-level n-grams captures lexical and morphological traits, hence appropriate for informal and noisy writing. This method adapts these embeddings to the particular features of Hinglish by fine-tuning both models on domain-specific Hinglish data. Dynamic combination of mBERT and FastText based on attention-based fusion mechanism assigns best weights to each feature set according on their relevance to the task.

This guarantees a more whole knowledge of Hinglish text by considering both lexical and contextual subtleties.

The suggested structure rests on a multi-stage pipeline. First, an unlabelled Hinglish blog corpus is used to fine-tune mBERT and FastText embeddings so they would be more suited for Hinglish's language traits. Features then are retrieved from these fine-tuned embeddings for a labelled cyberbullying corpus. The attention-based fusion process then combines these aspects into a shared realm where their contributions are dynamically balanced. At last, a classifier analyses the fused embeddings to forecast whether a particular text has cyberbullying material. Cross-entropy loss drives the whole model from end to end.

Our efforts are three-fold: First, we build a strong framework using cutting-edge models to handle the complexity of Hinglish and show how well it detects cyberbullying. Second, we present an attention-based fusion technique that dynamically uses the strengths of both contextual and lexical embeddings, hence improving performance on code-switched language. Third, we open the path for more inclusive NLP systems by offering analysis of the general relevance of our methodology to additional low-resource, code-switched languages.

We assess our approach using a labelled Hinglish cyberbullying dataset and show notable progress over current techniques. Our results highlight the need of using several embeddings and dynamic fusion techniques to solve Hinglish's difficulties. This work highlights the promise of sophisticated NLP methods to produce safer and more inclusive digital environments, hence advancing the battle against cyberbullying in multilingual cultures.

We explore the technical aspects of the framework in the following sections: the creation of the attention-based fusion mechanism, the fine-tuning of mBERT and FastText, and the assessment technique. By means of this study, we hope to close the disparity in cyberbullying detection for code-switched languages.

## Literature Review

Natural Language Processing (NLP) has demonstrated enormous potential for addressing issues such as hate speech, fake news, harassment, cyberbullying, and abusive language. Many strategies have been suggested over years to handle these difficulties. Using Naive Bayes and SVM classifiers on a YouTube dataset, Dinkar *et al*. [2] created a system to identify cyberbullying content, particularly targeted on sexual and racial words. Their model performed 68.30% for racism and 80.20% for spotting material linked to sexuality. Reynolds *et al.* [3] similarly presented a social media optimisation method applied to the Formspring dataset, obtaining 78.5% accuracy in separating bullying from non-bullying posts.

Additional developments improved cyberbullying identification using embedding-based algorithms. For instance, Djuric *et al.* [4] used paragraph-to- vector-based models of comments on Yahoo social media to detect hate speech with an 80% accuracy. Building on this, Badjatiya *et al.* [5] achieved a high F1-score of 93% by suggesting a CNN and LSTM architecture for identifying racism, misogyny, and other harmful content on Twitter. Balakrishnan *et al.* [6] developed a cyberbullying detection system for Twitter users in 2020 including psychological elements including mood, emotion, and personality factors. Using a participant-vocabulary consistency-based objective function to identify bullying on sites like Twitter and Ask.fm, Raisi *et al.* [7] broadened the scope by addressing fast developing vocabularies in social interactions. Using their MySpace and Form spring datasets, Squicciarini *et al.* [8] also investigated user demographics and social network aspects to examine peer pressure and social dynamics.

Transformer-based models especially BERT (Bidirectional Encoder Representations from Transformers) have lately been embraced rather extensively for abusive language detection. For sentence- and token-level tasks, BERT is a state-of- the-art solution because of its bidirectional context leveraging capabilities. Using news items, Aggarwal *et al.* [9] showed how strong fine-tuned BERT is at spotting bogus news. Fine-tuned on biomedical literature, domain-specific adaptations such as BioBERT [10] have surpassed previous models in tasks including information extraction and text mining. Stacked-DeBERT [11] used a customised encoding technique to improve classification in skewed and partial samples, hence highlighting BERT's adaptability.

BERT has also been used somewhat extensively to identify hate speech and poisonous remarks. To use pre-trained BERT's syntactic and contextual awareness for hate speech identification, Mozafari *et al.* [12] for example used a transfer learning method. In a similar vein, Pavlopoulos *et al.* [13] found objectionable language in talks across several social networking sites using a perspective-based fine-tuning technique. Offering strong solutions for several NLP chores, including offensive language detection, these transformer-based models have transformed the industry and opened the path for further developments.

Notwithstanding these developments, there are still difficulties especially in managing sophisticated, code-switched languages such as Hinglish. The special qualities of these languages marked by informal tone, code-mixing, and fast expanding vocabulary demand creative solutions combining contextual knowledge with lexical richness. This emphasises the requirement of fresh models combining modern embeddings with attention-based fusion to handle the complexity of low-resource and multilingual languages.

Al-Ajlan and Ykhlef [14] created a detection approach for 20,000 tweets that used deep convolutional neural networks (DCNN) for classification after substantial data cleaning and labelling. Their results were not encouraging, nonetheless, and imply that a bigger, multilingual sample might increase performance. Likewise, Banerjee *et al.* [15] used DCNN with GloVe embeddings on 69,874 tweets, obtaining a 93.7% accuracy, hence underscoring the possibility of applying this method to multilingual settings like Hindi-English. Emphasising the Wiki-Detox dataset, Wulczyn *et al.* [16] showed that their classifier produced AUC and Spearman correlation findings on par with human annotators, hence highlighting its possible for cyberbullying detection. With majority voting, the paper in [17] presented a reinforcement learning-based framework incorporating several NLP approaches and human-like behavioural patterns. Combining cross-entropy loss and Adam optimisation to exceed conventional approaches, Mahat*et al.* [18] used LSTM layers for cyberbullying detection across platforms including Twitter and Formspring. Yadav *et al.* [19] finally examined deep learning methods including CNN, RNN, LSTM, and BERT, finding that BERT's capacity to grasp phrase associations and contextual nuances led to better performance in weighted and unweighted classification tests.

## Methodology

The proposed approach aims to detect cyberbullying in Hinglish text by combining fine-tuned mBERT and FastText embeddings with an attention-based fusion mechanism. Hinglish, being a code-mixed language, presents unique challenges in natural language processing due to its informal nature, creative spellings, and blending of Hindi and English. To address these challenges, the methodology consists of the following steps: (1) dataset preparation and preprocessing, (2) fine-tuning FastText embeddings, (3) fine-tuning mBERT using masked language modeling, (4) attention-based feature fusion, and (5) classification using a supervised learning model. The figure 1 shows the proposed framework.

### A. Dataset Preparation and Preprocessing

Two datasets were used for this study: (1) a Hinglish blog corpus for domain adaptation and (2) a labeled Hinglish cyberbullying dataset for supervised learning. The Hinglish blog corpus, collected from various online platforms, contained informal text representing the linguistic diversity of Hinglish. The labeled dataset included annotated samples indicating whether a given sentence exhibited cyberbullying behavior.

Preprocessing was performed to clean and standardize the datasets. Text was converted to lowercase, repeated characters were normalized, and hashtags, emojis, and punctuation were retained where they contributed to meaning. Stop words were preserved, as they play an integral role in Hinglish syntax. For tokenization, FastText used word-level tokens, while mBERT employed subword tokenization to handle out-of-vocabulary words effectively.

### B. Fine-Tuning FastText for Hinglish

FastText embeddings were fine-tuned on the Hinglish blog corpus to capture domain-specific linguistic patterns. Using the skip-gram model, FastText learned to predict the context words of a given target word, leveraging its subword-level modeling capability. Words were represented as a sum of their n-gram embeddings, allowing the model to generalize across variations in spelling and transliteration.

Key hyper parameters included a vector dimension of 300, a window size of 5, and 5 negative samples per positive pair. Subword units were defined as character n-grams of sizes 3 to 6, enabling the model to capture morphological patterns in Hinglish text. The fine-tuning process minimized the skip-gram loss using stochastic gradient descent, updating embeddings iteratively to align with the Hinglish domain.

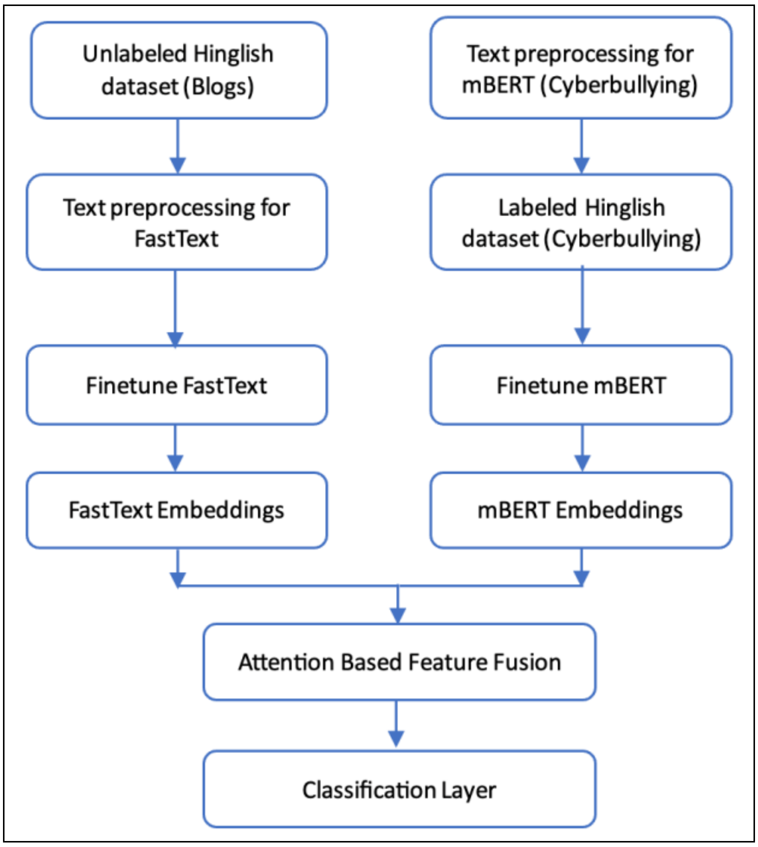
Unlabeled Hinglish dataset (Blogs)
Text preprocessing for FastText
Finetune FastText
FastText Embeddings
Text preprocessing for mBERT (Cyberbullying)
Labeled Hinglish dataset (Cyberbullying)
Finetune mBERT
mBERT Embeddings
Attention Based Feature Fusion
Classification Layer

**Fig 1:** Proposed Framework

### C. Fine-Tuning mBERT for Hinglish

To adapt mBERT to Hinglish, the model was fine-tuned using masked language modeling (MLM) on the Hinglish blog corpus. In MLM, a subset of tokens in each sentence was randomly masked, and the model was trained to predict these masked tokens based on the surrounding context. This process allowed mBERT to learn the syntactic and semantic nuances of Hinglish, including code-mixed structures and informal expressions.

The fine-tuning process involved tokenizing sentences into subwords and randomly masking 15% of the tokens. The model's parameters were updated by minimizing the cross-entropy loss between the predicted and actual masked tokens. Key hyperparameters included a learning rate of 2×10−52 \times 10^{-5}2×10−5, a batch size of 16, and 3 epochs of training. The fine-tuned mBERT provided contextualized embeddings that captured deeper semantic relationships in Hinglish text.

### D. Attention-Based Feature Fusion

The embeddings generated by mBERT and FastText were combined using an attention-based fusion mechanism. This approach dynamically assigned weights to the embeddings, enabling the model to determine the relative importance of contextualized and static representations for each input sentence. The fusion mechanism consisted of three steps: feature projection, attention score calculation, and weighted combination.

First, mBERT and FastText embeddings were projected into a shared dimensional space using linear transformations. Attention scores were computed for each embedding type using a softmax function, ensuring that their contributions were proportional to their relevance for the given input. Finally, the embeddings were combined into a single feature vector by taking a weighted sum of the transformed embeddings. This process ensured that the model leveraged the strengths of both embeddings while minimizing redundancy and conflicts.

### E. Classification

The fused feature vector was passed through a dense neural network for classification. The network consisted of a fully connected layer followed by a softmax function to compute the probabilities of the input belonging to each class. For binary classification (cyberbullying or non-cyberbullying), the cross-entropy loss was used as the objective function.

The classification model was trained on the labeled cyberbullying dataset, with the parameters of the attention mechanism and dense layer being optimized during training. The Adam optimizer was employed with a learning rate of 3×$10^{-5}$. Evaluation was performed using metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the model's performance.

This methodology integrates fine-tuned static and contextual embeddings with an attention-based fusion mechanism, enabling robust cyberbullying detection in Hinglish. By leveraging domain-specific embeddings and dynamic weighting, the proposed approach addresses the linguistic challenges posed by Hinglish and achieves improved performance in detecting harmful content.

### Proposed algorithm for Hinglish Cyberbullying Detection using mBERT and Fine-Tuned FastText with Attention-Based Fusion

**Inputs**

1. Unlabelled Hinglish blog dataset: $\mathcal{D}_{\text{Hinglish}} = \{x_1, x_2, \ldots, x_N\}$
2. Labeled Hinglish cyberbullying dataset: $D_{cyberbullying} = \{(x_i, y_i)\}$, where $y_i \in \{0,1\}$.
3. Pre-trained mBERT and pre-trained FastText embeddings.

**Outputs**

- Predicted label $y_i \in$ {0, 1} for each sentence $x_i$.

**Algorithm Steps**

**Step 1: Fine-Tune FastText on Hinglish Blogs**

Input Hinglish blog dataset $D_{\text{Hinglish}}$. Tokenize Hinglish text into words: $W(x_i) = \{w_1, w_2, \ldots, w_n\}$ and Normalize text (e.g., remove special characters, handle slang). Train FastText on the corpus using skip-gram or CBOW objectives. Output is Fine-tuned FastText embeddings.

$$L_{\text{fasttext}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|V|} p(w_j | w_{\text{context}}) \log q\left(w_j | w_{\text{context}}\right)$$

**Step 2: Fine-Tune mBERT on Hinglish Blogs**

Input Hinglish blog dataset $\mathcal{D}_{\text{Hinglish}}$. Tokenize sentences into sub words using mBERT tokenizer: $T(x_i) = \{t_1, t_2, \ldots, t_L\}$. Randomly mask 15% of tokens and fine-tune mBERT using the MLM objective, the output is Hinglish-adapted mBERT model.

$$L_{MLM} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{|V|} y_j^k \log \hat{t}_j^k$$

**Step 3: Feature Extraction for Cyberbullying Dataset**

Input: Labeled dataset $D_{cyberbbullying} = \{(x_i, y_i)\}$ Extract mBERT Features, by Tokenizing $T(x_i) = \{t_1, t_2, \ldots, t_L\}$ and passing through fine-tuned mBERT to get the contextualized [CLS] token embedding: $f_{mbert} x_i = \text{mBERT}(T(x_i))_{\text{CLS}}$ , to extract FastText Features Tokenize $x_i$ into words: $W(x_i) = \{w_1, w_2, \ldots, w_n\}$, and compute mean-pooled FastText embeddings:

$$f_{fasttext} x_i = \frac{1}{n}\sum_{j=1}^{n} f_{fasttext} x_i$$

**Step 4: Attention-Based Feature Fusion**

Normalize Embeddings $f_{mbert} x_i$ and $f_{fasttext} x_i$to unit length

$$f_{mbert} x_i = \frac{f_{mbert} x_i}{||f_{mbert} x_i||}$$

$$f_{fasttext} x_i = \frac{f_{fasttext} x_i}{||f_{fasttext} x_i||}$$

Then Transform Features into a Shared Space, project embeddings into a shared space $d_s$

$$z_{mbert} x_i = w_{mbert} \cdot f_{mbert} x_i + b_{mbert}$$

$$z_{fasttext} x_i = w_{fasttext} \cdot f_{fasttext} x_i + b_{fasttext}$$

then compute attention scores and compute attention weights for mBERT and FastText features,

$$\alpha_{\text{mbert}} = \text{softmax}(W_\alpha \cdot z_{mbert} x_i),$$

$$\alpha_{\text{fasttext}} = \text{softmax}\left(W_\alpha \cdot z_{fasttext} x_i\right)$$

and Normalize Attention Weights to ensure the attention weights sum to 1

$$\beta_{\text{mBERT}} = \frac{\alpha_{\text{mBERT}}}{\alpha_{\text{mBERT}} + \alpha_{\text{FastText}}}.$$

$$\beta_{\mathrm{FastText}} = \frac{\alpha_{\mathrm{FastText}}}{\alpha_{\mathrm{mBERT}} + \alpha_{\mathrm{FastText}}}$$

$$f_{combined}\, x_i = \beta_{\mathrm{mBERT}} \cdot z_{mbert}\, x_{i+}\beta_{\mathrm{FastText}} \cdot\ z_{fasttext}\, x_i$$

**Step 5: Classification**

Pass the fused feature vector through a classifier: $z^{(i)} = W_{\mathrm{classifier}} \cdot f_{combined}\, x_i + b_{\mathrm{classifier}}$, to predict compute probabilities using $\hat{y}_i = \mathrm{softmax}(z^{(i)})$, Predicted label: $\hat{y}_i = \arg\max_k \widehat{y_i^k}$

**Results and Discussion**

The experiments were conducted on a high-performance machine equipped with an Intel i9 9900K processor, 32 GB of RAM, and an NVIDIA 3080 GPU. This hardware configuration enabled efficient processing of large datasets and fine-tuning computationally intensive models like mBERT. The proposed models and baseline approaches were implemented using the Hugging Face Transformers library for mBERT, the FastText library for static embeddings, and PyTorch as the backend for training and evaluation. This setup ensured reproducibility and scalability for the experiments.

To evaluate the effectiveness of the proposed fusion framework, we compared its performance with several baseline approaches. The first baseline was a pre-trained mBERT model, used without domain adaptation or fine-tuning. This setup directly utilized the [CLS] token embeddings from mBERT for binary classification, providing insight into how well a general-purpose multilingual transformer performs on Hinglish text.

The second set of baselines utilized traditional machine learning models combined with static embeddings from FastText. These models included Logistic Regression, Support Vector Machine (SVM), and Random Forest. FastText embeddings were pre-trained on the Hinglish blog corpus, ensuring they captured the subword-level nuances of Hinglish. These traditional models offered a comparison point to assess the utility of static embeddings versus contextualized representations, the results are shown in Table 1.

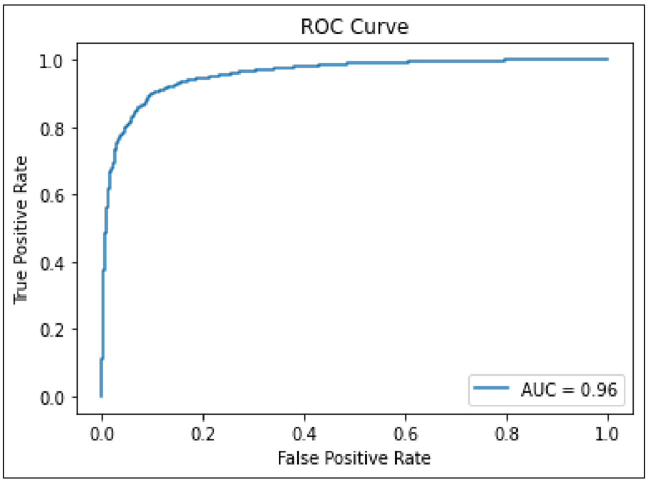

**Fig 2:** ROC curve for fusion model

Performance was evaluated using a comprehensive set of metrics to ensure a well-rounded analysis of each model's effectiveness. These included accuracy, which measures the proportion of correctly classified samples, and precision, which evaluates the proportion of true positive predictions among all positive predictions. Recall quantified the model's sensitivity to identifying true positives, while the F1-Score, the harmonic mean of precision and recall, provided a balanced metric. Finally, AUC (Area under the curve) assessed the model's ability to distinguish between positive and negative classes, offering insight into its overall discriminative power. The AUC curve for the fusion framework is shown in Figure 2, Figure 3 shows the precision recall curve, Figure 4 shows the confusion matrix of the fusion model and Figure 5 shows the training and validation loss curve.

**Table 1:** Results of model performance

| Model | Accuracy | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|
| mBERT Pre-trained | 0.477866 | 0.4904 | 0.8254 | 0.61525 | 0.439 |
| LR (FastText) | 0.88135 | 0.87862 | 0.8880 | 0.8833 | 0.946 |
| SVM (FastText) | 0.90950 | 0.91127 | 0.9096 | 0.91045 | 0.966 |
| RF (FastText) | 0.86016 | 0.85496 | 0.8713 | 0.86307 | 0.935 |
| Fusion model | 0.96579 | 0.98384 | 0.9887 | 0.96591 | 0.993 |

The pre-trained mBERT baseline exhibited suboptimal performance, with an accuracy of 0.4779, precision of 0.4904, recall of 0.8254, F1-score of 0.6153, and AUC of 0.4398. While mBERT achieved relatively high recall, indicating sensitivity to positive instances, its low precision and AUC demonstrated poor overall discrimination. This highlighted the challenges of applying a general-purpose multilingual transformer to Hinglish text without domain-specific fine-tuning.

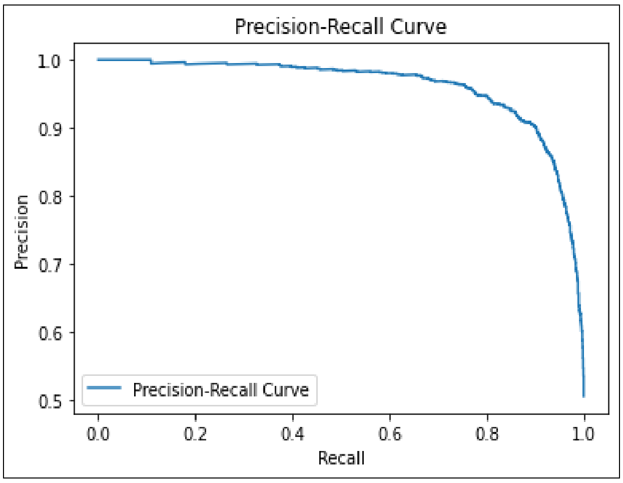

**Fig 3:** Precison recall curve for fusion model

In contrast, traditional machine learning models using fine-tuned FastText embeddings performed significantly better. Logistic Regression achieved an accuracy of 0.8814, precision of 0.8786, recall of 0.8881, F1-score of 0.8833, and AUC of 0.9466. SVM outperformed Logistic Regression with an accuracy of 0.9095, precision of 0.9113, recall of 0.9096, F1-score of 0.9105, and AUC of 0.9668, reflecting its ability to effectively utilize static embeddings for classification. Random Forest performed slightly lower, with an accuracy of 0.8602, precision of 0.8550, recall of 0.8713, F1-score of 0.8631, and AUC of 0.9359. These results underscore the effectiveness of traditional models combined with domain-adapted static embeddings.

The proposed fusion framework, integrating fine-tuned mBERT and FastText embeddings with an attention-based mechanism, achieved the highest performance across all metrics. It obtained an accuracy of 0.9658, precision of 0.9838, recall of 0.9881, F1-score of 0.9659, and AUC of 0.9935. The attention mechanism dynamically balanced the contributions of contextualized mBERT embeddings and static FastText embeddings, enabling the model to leverage the complementary strengths of both representations. This approach proved particularly effective in capturing the nuances of Hinglish, including code-mixed structures, creative spellings, and informal expressions, which are often challenging for individual models to handle.

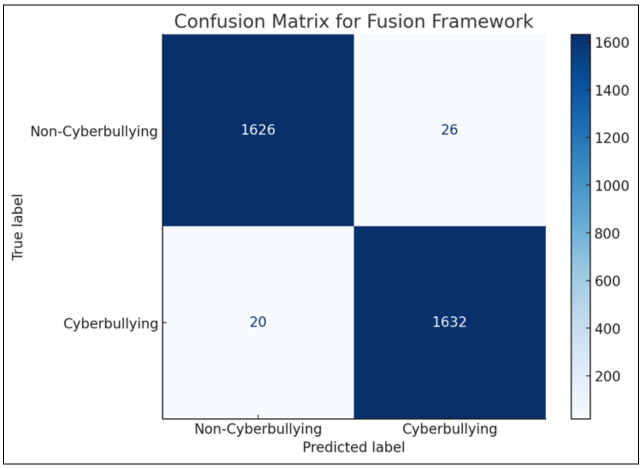

**Fig 4:** Confusion matrix for fusion model

The results of our experiments provide several key insights into the challenges and opportunities of cyberbullying detection in Hinglish. Hinglish, as a code-mixed language, introduces unique complexities due to its informal nature, creative spellings, and transliteration inconsistencies. Our findings highlight the importance of tailoring language models to these nuances and demonstrate the effectiveness of combining static and contextual embeddings using an attention-based fusion mechanism.

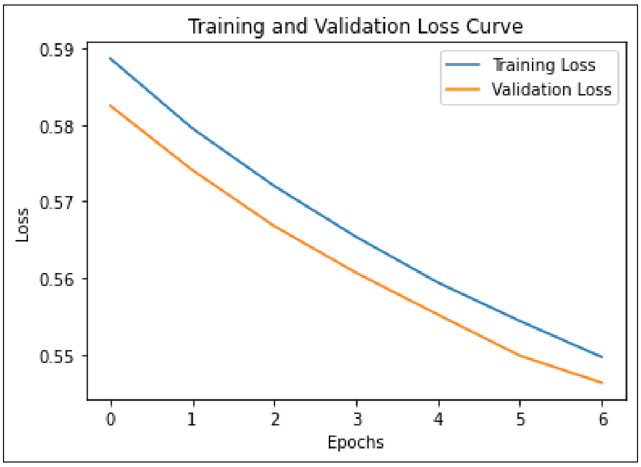

**Fig 5:** Training and Validation loss curve

The proposed fusion framework significantly outperformed all baseline models, achieving the highest performance across accuracy, precision, recall, F1-score, and AUC. This success can be attributed to the complementary strengths of the two embedding types. mBERT, with its contextualized embeddings, excelled at capturing sentence-level semantics and long-range dependencies, while Fast Text's sub word-level modelling effectively addressed transliteration and spelling variations common in Hinglish. By combining these embeddings through an attention-based mechanism, the model dynamically adjusted the contributions of each representation, focusing on the most relevant features for each input. This dynamic weighting enabled the framework to robustly handle diverse linguistic constructs in Hinglish.

The high precision and recall achieved by the fusion framework are particularly noteworthy. Precision reflects the model's ability to avoid false positives, ensuring that non-cyberbullying instances are not incorrectly flagged. This is critical for practical applications, where false accusations can have serious consequences. Similarly, high recall ensures that genuine instances of cyberbullying are not overlooked, providing comprehensive coverage of harmful content. The exceptional AUC score further underscores the model's ability to distinguish between cyberbullying and non-cyberbullying text, indicating its reliability in real-world scenarios.

The performance of pre-trained mBERT without domain adaptation was notably poor, highlighting the limitations of applying general-purpose multilingual models to Hinglish. While mBERT exhibited high recall, its low precision and AUC indicated a tendency to overgeneralize, resulting in many false positives. This reinforces the need for domain-specific fine-tuning, particularly for code-mixed languages that deviate significantly from standard language structures.

Traditional machine learning models with FastText embeddings provided strong baselines, with SVM achieving the best performance among them. These results demonstrate the utility of fine-tuned static embeddings for tasks involving transliteration and informal language. However, the performance of these models plateaued, as they lacked the ability to capture deeper contextual relationships within the text. This limitation was effectively addressed by the fusion framework, which combined the strengths of both static and contextual embeddings.

Despite its strong performance, the fusion framework is not without limitations. The computational complexity of combining mBERT and FastText embeddings, coupled with the attention mechanism, requires substantial resources. While this was mitigated in our experiments by using high-performance hardware, it may pose challenges for deployment in resource-constrained environments. Future work could explore strategies to optimize the computational efficiency of the framework, such as knowledge distillation or model pruning.

Another potential challenge lies in the inherent biases of pre-trained models. Both mBERT and FastText are trained on general corpora that may underrepresent or misrepresent certain linguistic patterns or cultural nuances specific to Hinglish. Although fine-tuning mitigates some of these biases, further steps, such as debiasing techniques or augmenting training data with more diverse examples, could enhance fairness and inclusivity.

The findings of this study have significant implications for real-world applications of cyberbullying detection in Hinglish and other code-mixed languages. The proposed fusion framework sets a new benchmark for performance, demonstrating the value of leveraging complementary embedding types. However, practical deployment of such systems must consider ethical implications, particularly in terms of false positives and the potential misuse of automated detection tools. Future research should prioritize developing explainable models that provide interpretable insights into their predictions, enabling better understanding and trust among users.

Additionally, extending the proposed methodology to other domains, such as hate speech detection or sentiment analysis in code-mixed languages, presents an exciting avenue for future work. Exploring cross-lingual transfer learning and domain adaptation techniques could further enhance the generalizability and robustness of the framework, making it applicable to a wider range of languages and contexts.

In summary, this study demonstrates the efficacy of combining fine-tuned mBERT and FastText embeddings with attention-based fusion for cyberbullying detection in Hinglish. By addressing the unique challenges of this code-mixed language, the proposed framework represents a significant step forward in the development of robust and reliable NLP systems for multilingual and informal text.

## Conclusion

This study proposed a hybrid framework for cyberbullying detection in Hinglish text, integrating fine-tuned mBERT embeddings with FastText embeddings using an attention-based fusion mechanism. Hinglish, being a code-mixed language with informal expressions and transliteration inconsistencies, presents unique challenges for natural language processing tasks. The experimental results demonstrated that the proposed approach effectively addresses these challenges, achieving superior performance compared to both pre-trained mBERT and traditional machine learning models with static embeddings.

The pre-trained mBERT baseline struggled to generalize to Hinglish text, highlighting the limitations of applying general-purpose multilingual models to code-mixed languages without domain adaptation. While traditional machine learning models with FastText embeddings provided strong baselines, their reliance on static embeddings limited their ability to capture deep contextual relationships. These findings emphasize the need for models that combine the strengths of both static and contextualized representations.

The proposed fusion framework achieved remarkable results, with an accuracy of 96.58%, precision of 98.38%, recall of 98.81%, F1-score of 96.59%, and an AUC of 99.35%. The attention-based mechanism dynamically balanced the contributions of mBERT and FastText features, enabling the model to focus on the most relevant aspects of each representation. This approach effectively captured the linguistic intricacies of Hinglish, including creative spellings, code-mixed structures, and informal syntax, resulting in a robust and reliable system for detecting cyberbullying.

The results of this study demonstrate the potential of hybrid embedding techniques in addressing complex NLP challenges in low-resource and code-mixed languages. By leveraging the complementary strengths of mBERT and FastText, the proposed framework sets a new benchmark for cyberbullying detection in Hinglish. Beyond this task, the methodology can be extended to other applications and languages, offering a flexible and powerful solution for multilingual and code-mixed text processing.

Ethical considerations, such as addressing potential biases in the data and ensuring fairness in predictions, remain critical for deploying such systems in real-world applications. Overall, this study underscores the importance of hybrid models in advancing NLP for diverse and complex linguistic landscapes.

## References

1. Kowalski RM, Limber SP. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. J Adolesc Health. 2013 Jul;53(1 Suppl):S13-20. DOI: 10.1016/j.jadohealth.2012.09.018.
2. Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying. In: Proceedings of the International AAAI Conference on Web and Social Media; 2011 May 17-20; Barcelona, Spain. Palo Alto (CA): AAAI Press, 2011, p. 1-6. DOI: 10.1609/icwsm.v5i3.14209.
3. Reynolds K, Kontostathis A, Edwards L. Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops; 2011 Dec 12-15; Honolulu, HI, USA. IEEE; 2011. p. 241-244. DOI: 10.1109/ICMLA.2011.152.
4. Hate speech detection with comment embeddings. Proceedings of the 24th International Conference on World Wide Web; 2015 May 18-22; Florence, Italy. Available from: https://dl.acm.org/doi/10.1145/2740908.2742760.
5. Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion; 2017 Apr 3-7; Perth, Australia. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, p. 759-760. DOI: 10.1145/3041021.3054223.
6. Balakrishnan V, Khan S, Arabnia HR. Improving cyberbullying detection using Twitter users' psychological features and machine learning. Comput Secur. 2020 Mar;90:101710. DOI: 10.1016/j.cose.2019.101710.
7. Raisi E, Huang B. Cyberbullying identification using participant-vocabulary consistency. ArXiv. 2016 Jun. Available from: https://www.semanticscholar.org/paper/Cyberbullying-Identification-Using-Consistency-Raisi-Huang/8813d9cf19e201bf81e1d919a098e7f5921954e3.
8. Squicciarini A, Rajtmajer S, Liu Y, Griffin C. Identification and characterization of cyberbullying dynamics in an online social network. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2015 Aug 24-27; Paris, France. IEEE, 2015, p. 280-285. DOI: 10.1145/2808797.2809398.
9. Aggarwal A, Chauhan A, Kumar D, Verma S, Mittal M. Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Trans Scalable Inf Syst. 2020 Apr;7(27):e163973. DOI: 10.4108/eai.13-7-2018.163973.
10. Lee J, Yoon W, Kim S, Kim D, So CH, Lee H. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb;36(4):1234-1240. DOI: 10.1093/bioinformatics/btz682.

11. Cunha G, Lee M. Stacked DeBERT: all attention in incomplete data for text classification. Neural Networks. 2021 Apr;136:87-96. DOI: 10.1016/j.neunet.2020.12.018.
12. Mozafari M, Farahbakhsh R, Crespi N. A BERT-based transfer learning approach for hate speech detection in online social media. In: Cherifi H, Gaito S, Mendes JF, Moro E, Rocha LM, editors. Complex networks and their applications VIII. Studies in Computational Intelligence, vol. 886. Cham: Springer International Publishing, 2020, p. 928-940. DOI: 10.1007/978-3-030-36687-2_77.
13. Pavlopoulos J, Thain N, Dixon L, Androutsopoulos I. ConvAI at SemEval-2019 Task 6: offensive language identification and categorization with Perspective and BERT. In: May J, Shutova E, Herbelot A, Zhu X, Apidianaki M, Mohammad SM, editors. Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), 2019 Jun, p. 15-16, Minneapolis, Minnesota, USA. Stroudsburg (PA): Association for Computational Linguistics, 2019, p. 571-576. DOI: 10.18653/v1/S19-2102.
14. Al-Ajlan MA, Ykhlef M. Optimized Twitter cyberbullying detection based on deep learning. In: 2018 21st Saudi Computer Society National Computer Conference (NCC), 2018 Apr, p. 23-25, Jeddah, Saudi Arabia. IEEE, 2018, p. 1-5. DOI: 10.1109/NCG.2018.8593146.
15. Banerjee V, Telavane J, Gaikwad P, Vartak P. Detection of cyberbullying using deep neural network. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS); 2019 Mar 15-17; Coimbatore, India. IEEE, 2019, p. 604-607. DOI: 10.1109/ICACCS.2019.8728378.
16. Wulczyn E, Thain N, Dixon L. Ex Machina: personal attacks seen at scale. arXiv. 2017 Feb 25. Available from: https://arxiv.org/abs/1610.08914.
17. Aind AT, Ramnaney A, Sethia D. Q-Bully: a reinforcement learning based cyberbullying detection framework. In: 2020 International Conference for Emerging Technology (INCET), 2020 Jun, p. 26-28, Vellore, India. IEEE; 2020, p. 1-6. DOI: 10.1109/INCET49848.2020.9154092.
18. Mahat M. Detecting cyberbullying across multiple social media platforms using deep learning. In: 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE); 2021 Mar 19-21; Ghaziabad, India. IEEE, 2021, p. 299-301. DOI: 10.1109/ICACITE51222.2021.9404736.
19. Yadav Y, Bajaj P, Gupta RK, Sinha R. A comparative study of deep learning methods for hate speech and offensive language detection in textual data. In: 2021 IEEE 18th India Council International Conference (INDICON); 2021 Dec 17-19; Delhi, India. IEEE, 2021, p. 1-6. DOI: 10.1109/INDICON52576.2021.9691704.