Detection of Misogyny in Hindi Code-Mixed Texts by BiGRU with Bahdanau Attention using ByT5 Embeddings

S. Karishma, and V. Akila

Original scientific article

***Abstract*—Social media platforms became the hub for conveying messages and responses to current events, but contain more harmful aspects by influencing negative stereotypes, spreading false information, and enabling misogyny in some scenarios. Detecting misogynistic language in social media is challenging for code-mixed languages due to demographic variations, transliteration, and noisy texts. The model has been evaluated on a Hindi-English code-mixed dataset of misogynistic comments. We proposed a hybrid misogyny classification model that combines byte-level ByT5 encoder embeddings with a Bidirectional Gated Recurrent Unit (BiGRU) augmented by the Bahdanau attention mechanism. ByT5 produces robust, subword-agnostic representations that reduce sensitivity to spelling variations and code switching. The BiGRU captures contextual sequential patterns and bidirectional dependencies, while attention emphasizes the most indicative tokens of abusive intent. It is demonstrated that the proposed hybrid model outperforms recurrent neural networks with static and dynamic embeddings, producing more stable misogyny predictions in low-resource and noisy texts.**

***Index terms*—Bahdanau Attention Mechanism, Bidirectional Gated Recurrent Unit, ByT5 embedding, Code-mixed, Misogyny, Noisy text.**

I. Introduction

Nowadays, social media platforms have become the ultimate platform for communication, community, and forums, enabling the instant broadcast of messages, photos, and multimedia content. Cyberbullying and online harassment are predominant issues that harm users' online experiences and endanger their social, mental, and physical well-being. According to research, there is a significant correlation between online exposure to hate speech and abusive content and depressive symptoms, a poorer level of life satisfaction, and an increased likelihood of victimization. Thus, ensuring polite and safe online spaces becomes a top priority for social media users. The act of regret and prejudice against women is known as misogyny, propagated significantly through the Web, particularly via social media platforms targeting women [1].

Manuscript received November 19, 2025; revised December 22, 2025. Date of publication June 1, 2026. Date of current version June 1, 2026. The associate editor prof. Maja Braović has been coordinating the review of this manuscript and approved it for publication.

Authors are with the Department of Computer Science and Engineering, Puducherry Technological University, Puducherry, India (e-mails: 16karish@ptuniv.edu.in, akila@ptuniv.edu.in).

Digital Object Identifier (DOI): 10.24138/jcomss-2025-0240

Misogynistic content can be apparent in the form of insults, derogatory comments, or threats of violence, but it often takes the form of stereotypes, objectification, or the normalization of damaging gender norms. Misogyny spreads especially easily on social media, where anonymity or a lack of accountability sometimes lessens the perceived repercussions of abusive conduct [1].

While much of the existing computational research on misogyny, hateful content, or cyberbullying focuses more on globally adapted languages, especially English, there is increasing awareness of low-resource settings, such as languages or linguistic varieties that have limited annotated data, fewer computational tools, and less prior research. This is especially relevant, where writing one language in another language's script, known as code-mixing, is more common for multilingual people. In such contexts, abusive content often switches between languages (e.g., English and a regional language) or uses colloquial phrases in day-to-day life [1].

We hypothesize that integrating byte-level ByT5 embeddings with BiGRU and Bahdanau attention improves misogyny detection performance in Hindi code-mixed texts by enhancing robustness to orthographic noise and capturing contextual dependencies that are inadequately modeled by conventional word-level and transformer-only baselines.

The goal of the work described in this article is to detect and categorize misogynistic comments in low-resource, noisy, code-mixed scenarios to have a greater impact by mitigating negative stereotypes, spreading false information, and creating a safer space for women in the social media environment.

The contributions made in this article are:

1) Proposed hybrid architecture for misogyny detection using the Bidirectional Gated Recurrent Unit (BiGRU) model with ByT5 byte-level contextual embedding.
2) The hyperparameter tuning is applied using Optuna for finding the best fit in BiGRU hyperparameters.
3) Bahdanau attention is integrated for enhancing interpretability by preferencing influential words in predictions.

This article is organized in the following manner: Section II reviews the related literature. Dataset information and the methodology adopted are described in Section III. Section IV describes the implementation of the proposed framework, and Section V compares its results with those of existing approaches. The article concludes in Section VI.

II. Literature Survey

Recent studies have proposed that misogyny, hate speech, and offensive language detection increasingly rely on deep learning algorithms, transformer-based and hybrid architectures, especially in multilingual and low-resource language contexts, as described in Table I.

Singh et al. [2] assessed various machine learning and deep learning models and presented a high-quality annotated dataset for identifying sexist sentiments in YouTube comments. Their research showed that transformer-based models outperform conventional classifiers, underscoring the importance of contextual embeddings. Saumya et al. [3] investigated offensive-language filtering in multilingual social media using deep learning architectures, such as CNNs and LSTMs. The fine-tuned BERT outperformed other models. Nevertheless, the study relies on word-level tokenization and does not explore robustness to spelling variations or transliteration commonly observed in social media. Neog et al. [4] presented a hybrid deep learning model that combines convolutional and recurrent layers to detect offensive comments in Assamese, resulting in the wellness of hybrid systems working for languages with limited resources. However, the model's inability to capture implicit toxicity and contextual dependencies stems from its reliance on static embeddings.

Khanduja et al. [5] developed a Telugu hate speech corpus and evaluated transformer-based models, reporting substantial performance gains with multilingual transformers. Similarly, Rajalakshmi et al. [6] proposed HOTTEST, a transformer-based framework with enhanced stemming for Tamil offensive content detection, achieving improved accuracy. Despite these advances, these studies primarily employ token-based transformers and do not explicitly address orthographic noise or code-mixed scenarios. Roy et al. [7] proposed a deep ensemble framework for hate speech identification in Dravidian languages, showing that ensemble learning enhances performance and resilience.

Ali et al. [8] investigated the transfer learning of pre-trained language models across domains for Twitter hate speech detection. Chakravarthi et al. [9] outperformed baseline models by identifying objectionable language in Dravidian languages using MPNet embeddings combined with CNN model. Akhter et al. [10] proposed a hybrid machine learning model for Bengali cyberbullying detection, illustrating the benefits of combining multiple feature representations. While these approaches improve classification performance, they largely depend on word- or subword-level representations and are vulnerable to spelling noise.

Hashmi et al. [11] demonstrated that multilingual transformers improve sentiment prediction in code-mixed tweets, while Rosid et al. [12] employed a multi-head attention-based CNN-BiGRU architecture for sarcasm detection in Indonesian–English code-mixed text, highlighting the effectiveness of combining recurrent and attention mechanisms. Biradar et al. [13] proposed a Conv-LSTM Siamese network for Hindi–English code-mixed hate speech detection, achieving strong performance in low-resource settings. These studies emphasize the value of hybrid architectures but do not explore byte-level representations to handle extreme spelling variations.

TABLE I
Tabulated Literature Review

Ref. No.	*Dataset*	*Methodology*	*Results*
[2]	12,698 Hindi-English misogynistic comments	MNB, SVM, KNN, DT, LR with TF-IDF Embeddings CNN, LSTM with GLoVe embeddings and mBERT	Maximum weighted average F1-score using mBERT is 0.66 for subtask 1 and 0.65 for subtask 2, respectively.
[3]	42,560 social media comments from five languages (English, Hindi, German, Tamil, and Malayalam)	Word2Vec, GloVe, BERT, CNN, Bi-LSTM, Bi-LSTM-Attention, and fine-tuned BERT	Maximum macro-average scores of 0.79 for monolingual tasks and 0.86 for code-mixed tasks using finetuned BERT.
[4]	50,000 Assamese English code-mixed comments	CNN + BiLSTM with Sigmoid and Softmax activation function	Accuracy of 88.43% with sigmoid activation and 90.51% with the softmax activation function.
[5]	38,035 Telugu tweets	mBERT, DistilBERT, IndicBERT, NLLB, MuRIL, RNN+LSTM, XLM-RoBERTa, and IndicBART	Maximum accuracy of 98.2% using finetuned mBERT.
[6]	HASOC 2021 YouTube comments in the Tamil language	TF-IDF and pre-trained transformer models like BERT, XLM-RoBERTa, IndicBERT, mBERT, TaMillion, and MuRIL	Maximum F1-score of 84% and accuracy of 86% using MuRIL.
[7]	Malayalam and Tamil code-mixed datasets	LR, RF, MNB, XGB, SVM, DNN, CNN, LSTM. BERT, DstilBERT, XLM-RoBERTa, and MuRIL	Maximum weighted F1-score of 0.802 and 0.933 for Malayalam and Tamil code-mixed datasets using a deep ensemble framework.
[8]	10,526 Urdu tweets.	Machine learning algorithms and FastText Urdu word embeddings and mBERT, XLM-RoBERTa, and Distil-BERT	Achieved F1-scores of 0.68, 0.68, and 0.69 using mBERT, XLM-RoBERTa, and Distil-BERT, respectively.
[9]	DravidianCode Mixed dataset in Tamil, Malayalam, and Kannada	SVC, MNB, DT, RF, LGBM, EWDT, EOWDT, CNN, mBERT	Maximum weighted average F1-Score of 0.85, 0.98, and 0.76 for code-mixed Tamil, Malayalam, and Kannada, respectively.
[10]	44,001 Bengali cyberbullying text dataset	TFIDF + IHT + DT, RF, LR, MLP	Maximum accuracy of 98.57% in binary using LR and 98.82% in multilabel classification using MLP.

[11]	20735 Political Tweets - English, Roman Urdu, and mixed	Electra, code-mixed BERT (cm-BERT), and Multilingual Bidirectional and Auto-Regressive Transformers (mBART)	Maximum F1-Score of 0.73 using mBART.
[12]	Indonesian Tweets DS1 – 5000 News headlines DS2 - 28619	CNN with multi-head attention and BiGRU (MHA-CovBi) with other deep learning algorithms	Maximum accuracy of 94.6% and F1-Score of 94.38% using MHA-CovBi model.
[13]	Twitter comments of Hindi-English codemixed - 2914	Conv-LSTM-Based Siamese Network and other transformer models	Maximum accuracy of 72% using Conv-LSTM-Based Siamese Network model.

III. METHODOLOGY

A. Baseline Models

Static embedding has a constant vector value for each word, irrespective of the situation that appears. Examples of static embedding are Word2vec, GloVe, and FastText embeddings. The limitation of static embedding is that it cannot handle polysemy (words with multiple meanings). The change of word vectors depends on the context in which the word appears in contextual embedding to avoid false positives or negatives in misogyny detection. Polysemy is highly relevant in misogyny detection because misogynistic language often reuses common words in derogatory contexts. The contextual embeddings are performed by transformer-based models are suitable for handling polysemy.

For example,

Neutral: "The new item was bought from the store."
Misogynistic: "She's such an item." (objectification of women)

FastText embedding is a multilingual static embedding used for representing vectors from words created by Facebook's AI Research (FAIR) group, which provides pretrained embeddings for several languages that enhance the Word2Vec model by handling out-of-vocabulary (OOV) words more efficiently, expressed by the words as bags of character n-grams.

The polyglot variant of BERT (mBERT) is trained on the Wikipedia dataset in 104 languages. Developed for supporting transfer learning and cross-lingual understanding, which made it useful for zero-shot or few-shot learning tasks across multiple languages.

XLM-RoBERTa was created by Facebook AI. The multilingual transformer model was trained on 100 languages using the enormous CommonCrawl dataset. It is an improved version of RoBERTa that can perform better than mBERT.

Multilingual Representations for Indian Languages (MuRIL), developed by Google AI in 2021, are specifically trained for 17 Indian languages. It was built on mBERT but further pre-trained on Indian language corpora (Wikipedia, Common Crawl, parallel corpora, and transliterated text) that handle script variation, code-mixing, and transliteration better than generic multilingual models.

The existing neural network-based approaches with static word embeddings (e.g., GloVe, FastText) lack contextual sensitivity and struggle to capture polysemy. Transformer-based models such as BERT, RoBERTa, and XLM-R have substantially improved performance by leveraging contextualized representations but primarily rely on sequential dependencies, which are difficult to detect for subtle abusive intent, especially in short, informal, code-mixed texts.

By improving these gaps, the proposed work integrates byte-level ByT5 embeddings, which are inherently robust to spelling noise and code-mixed text, with a BiGRU and Bahdanau attention mechanism to capture both sequential dependencies and task-specific salient features. This design aims to address the limitations of state-of-the-art approaches, particularly in multilingual and noisy social media settings.

B. Proposed Methodology

The proposed architecture combines three powerful architectures: ByT5 embedding, BiGRU, and the Bahdanau attention mechanism, allowing the framework to capture rich contextual information to detect misogyny in Hindi code-mixed text. The architecture of the proposed model is displayed in Figure 1.

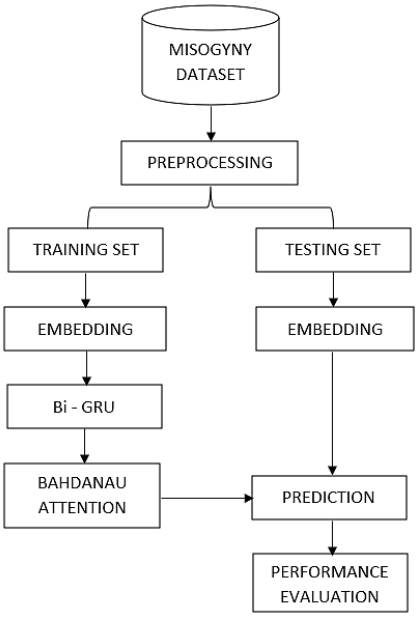

Fig. 1. Architecture diagram of the proposed model

A byte-level version of T5 (ByT5) embedding was developed by Google Research in 2021, which takes raw UTF-8 bytes as input, avoiding issues with tokenization and vocabulary. ByT5 works robustly across 100+ languages, eliminates dependency on language-specific tokenizers, and is inherently robust to spelling variations, orthographic noise, including multilingual and code-mixed text, especially where spelling variation or informal text is common on social media [14].

Algorithm 1: The BiGRU with the Bahdanau Attention Mechanism model with ByT5 embedding for misogyny classification

Input: Raw text dataset D = {X_i, Y_i} for i = 1 to N, where X_i is text, and Y_i is label

Output: Misogyny classification
1: Load dataset D from the CSV file
2: *for* each text X_i in D:
3: $X_i \leftarrow$ preprocess(X_i) # text cleaning, normalization
4: Encode labels Y_i using LabelEncoder
5: *for* each text X_i in D:
6: tokens ← ByT5_tokenizer(X_i)
7: embeddings ← ByT5_encoder(tokens)
8: $E_i \leftarrow$ mean(embeddings)
9: Convert E_i and Y_i into PyTorch tensors
10: Create TensorDataset(E_i, Y_i)
11: Prepare DataLoader for batching
12: Reshape embeddings to (batch_size, seq_len=1, embedding_dim)
13: Define BiGRU(input_dim, hidden_dim, bidirectional=True)
14: Define the Bahdanau attention mechanism
15: *for* each hidden state h_t in BiGRU output H = {h_1, h_2, ..., h_T}:
16: $e_{(t,i)} = v\_a^T * \tanh(W_1 * h_t + W_2 * s_i)$
17: $\alpha_{(t,i)} = \text{softmax}(e_{(t,i)})$
18: $c_t = \Sigma_i (\alpha_{\_(t,i)} * s_i)$
19: Concatenate [c_t ; h_t]
20: $h_{t'} = \tanh(W_c * [c_t ; h_t])$
21: Pass $h_{t'}$ through Fully Connected Layers with ReLU activation and dropout
22: Define search_space = {hidden_dim, learning_rate, dropout, batch_size, etc.}
23: *for* each trial in range(1, 10):
24: model ← BiGRU_Attention(trial_parameters)
25: Train model for 10 epochs on training set
26: Record validation loss
27: Select the best_trial ← trial with the minimum validation loss
28: Initialize model with best_trial hyperparameters
29: *for* epoch = 1 to 100:
30: Train model using RMSprop optimizer
31: Compute loss using categorical_cross_entropy
32: Monitor validation accuracy
33: *if* early_stopping_condition met:
34: Save best model weights
Break
35: Output final trained model
36: Predict labels on the test set using best_model
37: Compute performance metrics
38: Visualize model interpretability using LIME
39: Return final classification results

The Bidirectional Gated Recurrent Unit (BiGRU) overcomes the complex structure and higher computational cost of the LSTM by enhancing the LSTM architecture through the integration of a gating mechanism, thereby promoting more sequential tasks without compromising performance. BiGRU enhances contextual representation beyond transformer embeddings by explicitly modelling token-level temporal dependencies. BiGRU has two gates. The Reset gate determines how much prior hidden state should be forgotten, and the update gate determines how much new data should be updated into the secret state. The term 'bidirectional' means that the GRU propagates the sequence of inputs in both forward and backward directions, allowing it to handle both past and future tokens [15].

$$r_t = \sigma\,(W_r\,.\,[h_{t-1}, x_t] + b_r) \quad (1)$$

$$z_t = \sigma\,(W_z\,.\,[h_{t-1}, x_t] + b_z) \quad (2)$$

The attention mechanism is applied to inputs containing long sequences of words, assigning each word a weight based on its importance and assigning higher values to the most significant ones. The Bahdanau Attention, also known as an additive attention mechanism, drives the current state of the decoder, the previous state, and the current input through the neural network. Next, the attention weights are used to compute a weighted sum of the additional input features provided to the decoder, allowing it to focus on essential input factors when producing a single output. This mechanism improves interpretability and enables the network to emphasize abusive or misogynistic cues that may appear sparsely within long or noisy texts [16].

Alignment Score: Similarity measure between the decoder's current hidden state (what it's trying to generate) and each encoder's hidden state (each word/token in the input sequence).

$$e_{t,i} = v_a^T\,.\,\tanh\,(W_1 h_t + W_2 s_i) \quad (3)$$

Attention weights: Softmax over all i (input tokens) to get normalized attention scores.

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})} \quad (4)$$

Context Vector: The weighted total of the encoder's concealed state is the context vector.

$$c_t = \sum_i \alpha_{t,i}\,.\,s_i \quad (5)$$

Final output: The context vector c_t and decoder state h_t are combined to get a new hidden state.

$$h_t = \tanh\,(W_c\,[c_t; h_t]) \quad (6)$$

IV. Experimental Setup

A. Dataset Description

The dataset consists of 12,698 misogyny Hindi-English code-mixed YouTube comments collected and annotated by A. Singh et al. [2]. This dataset serves two subtasks: the first categorizing content into neutral, optimistic, or pessimistic sentiment, and the second further categorizing content into appreciation, suggestion, criticism, offensive, or unknown instances. The

dataset distribution for both subtasks is given in Table II, and the word cloud of the misogyny codemixed dataset is displayed in Figure 2.

TABLE II
DATASET DISTRIBUTION OF MISOGYNY HINDI CODE-MIXED DATASET

Tasks	*Label*	*Count*
Subtask 1	Optimistic	1629
	Pessimistic	3929
	Neutral	7140
Subtask 2	Appreciation	405
	Suggestion	639
	Criticism	2436
	Offensive	1493
	Unknown	7725

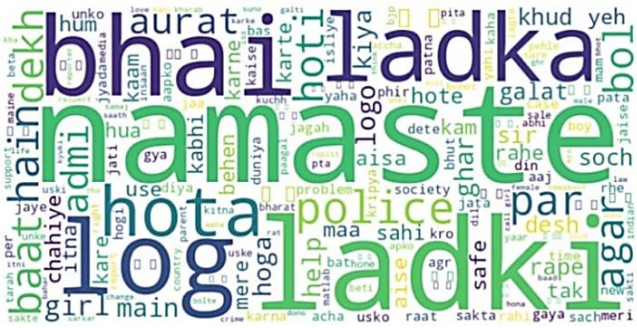

Fig. 2. Word cloud of the misogyny Hindi-English code-mixed dataset

B. Working

Since the local runtime connection resulted in frequent interruptions due to insufficient local computational resources for stable model training, the model was run on Google Colaboratory's cloud infrastructure, which provided an NVIDIA Tesla T4 GPU with approximately 16 GB of GPU memory. The implementation was carried out using Python 3.12.12 and PyTorch 2.9.0. The CSV file contains comments and labels, encoded using a label encoder. The dataset is split into a training and test set at a 75:25 ratio. The texts are tokenized with the ByT5 tokenizer, and the resulting tokens are passed through the ByT5 encoder. The mean for all token embeddings is calculated to produce a single fixed-length vector per text. The embeddings and labels are converted into PyTorch tensors for processing input in GRU. The BiGRU captures contextual sequential patterns and bidirectional dependencies, whereas the Bahdanau attention emphasizes tokens of most indicative of abusive intent. Hyperparameters are tuned with Optuna by conducting 10 trial combinations and training for 10 epochs to find the best hyperparameters for obtaining the final model. Table III describes the search space for hyperparameter tuning.

The final model yielded the best hyperparameters, such as a dropout rate of 0.3, a learning rate of 0.001, a batch size of 32, and a BiGRU size of 128. The configured model was trained for 100 epochs using the RMSprop optimizer, and the loss was validated using CrossEntropy with the softmax activation function for multiclass classification. The duration of the proposed model, trained after hyperparameter tuning for each epoch, is 65.14 s. The trained model was evaluated on the test set, and the proposed model's performance was assessed. The Local Interpretable Model-agnostic Explanations (LIME) visualizes the calculated probability for each word, which identifies the predicted label.

TABLE III
SEARCH SPACE FOR HYPERPARAMETER TUNING

Hyperparameters	*Range*	*Step Count*
Hidden Size of BiGRU	[64, 512]	64
Dropout Rate	[0.1, 0.5]	0.1
Learning Rate	[1e-4, 1e-2]	1e-2
Batch Size	[16, 64]	16

V. RESULTS AND DISCUSSION

The proposed BiGRU with Bahdanau attention and a byte-wise ByT5 encoder embedding achieves 63.27% accuracy on subtask 1 and 64.85% on subtask 2, outperforming static and contextual embeddings on the Hindi-English code-mixed dataset and handling low redundancy, code-mixed, and noisy texts well. The performance metrics of the proposed framework are evaluated against baselines for both subtasks of the Hindi-English code-mixed dataset and are tabulated in Tables IV and V. The graphs for the proposed framework with multiple models for both subtasks are shown in Figures 3 and 4.

TABLE IV
PERFORMANCE OF VARIOUS MODELS UNDER SUBTASK 1

Model	*Embedding*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
BiGRU	FastText	62.71	62.09	62.71	62.16
BiGRU	mBERT	62.45	61.15	62.45	61.49
BiGRU	XLM-RoBERTa	61.25	60.39	61.25	60.87
BiGRU	MuRIL	60.61	57.65	60.61	54.55
BiGRU+Att.	ByT5	63.27	61.54	63.27	61.55

TABLE V
PERFORMANCE OF VARIOUS MODELS UNDER SUBTASK 2

Model	*Embedding*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
BiGRU	FastText	64.78	64.59	64.78	64.94
BiGRU	mBERT	60.38	58.83	60.37	59.45
BiGRU	XLM-RoBERTa	61.54	60.56	61.54	60.95
BiGRU	MuRIL	62.74	58.75	62.74	59.31
BiGRU+Att.	ByT5	64.85	63.65	64.85	63.59

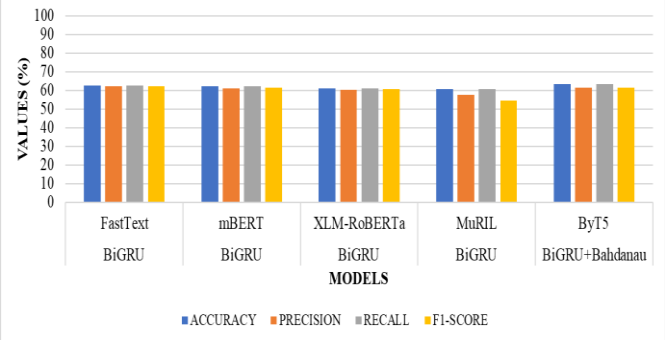

Fig. 3. Performance graph of the proposed model with other baselines in Subtask 1

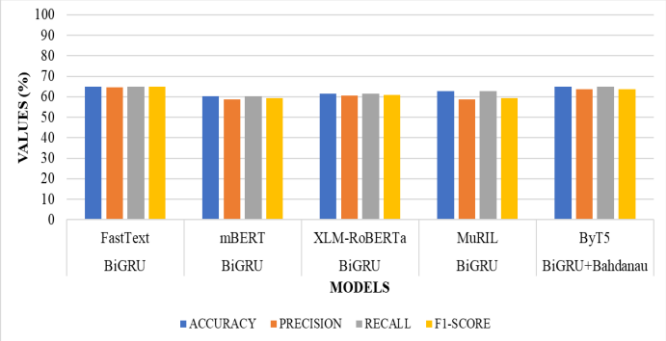

Fig. 4. Performance graph of the proposed model with other baselines in Subtask 2

Sample texts are tested using the proposed framework, and the prediction probability for each label is calculated, and the maximum value of the likelihood indicates how well the words are correctly predicted, with the actual label visualized using LIME, the figures 5, 6, and 7 for subtask 1, and the figures 8, 9, 10, 11, and 12 for subtask 2, respectively.

A. *Correctly classified sample texts using the proposed model to classify Subtask 1*

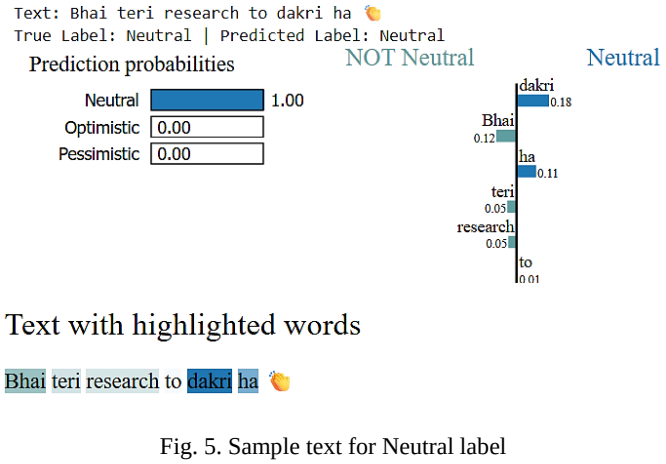

Fig. 5. Sample text for Neutral label

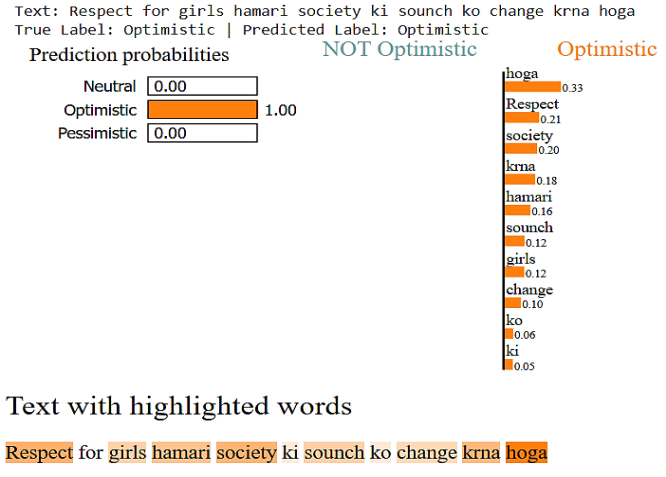

Fig. 6. Sample text for Optimistic label

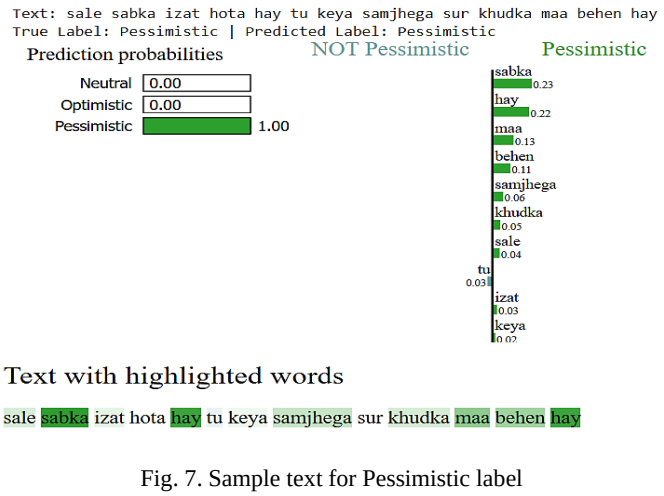

Fig. 7. Sample text for Pessimistic label

B. *Correctly classified sample texts using the proposed model to classify Subtask 2*

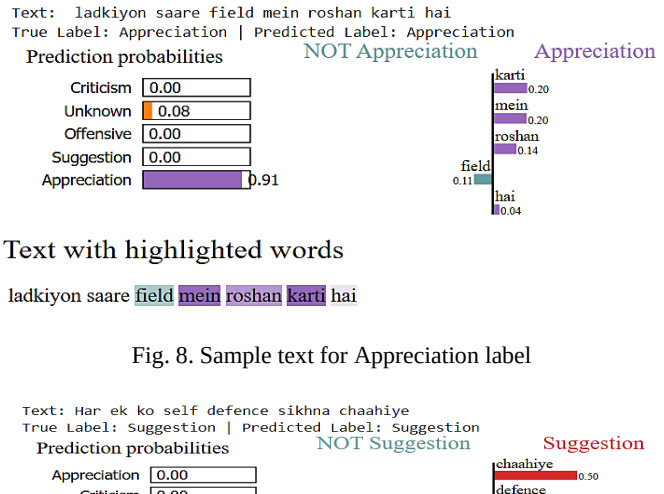

Fig. 8. Sample text for Appreciation label

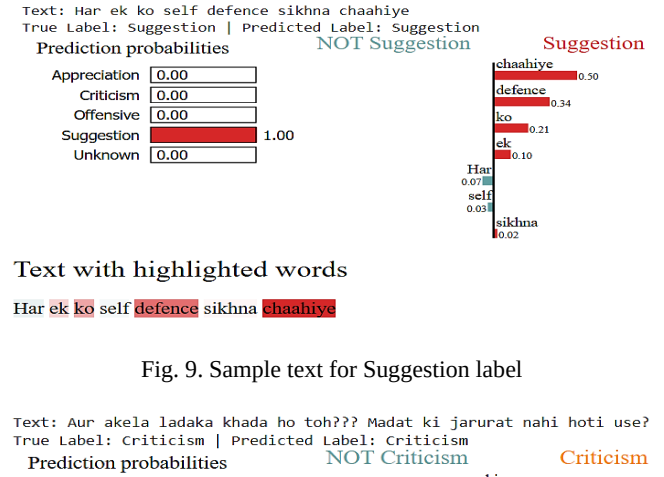

Fig. 9. Sample text for Suggestion label

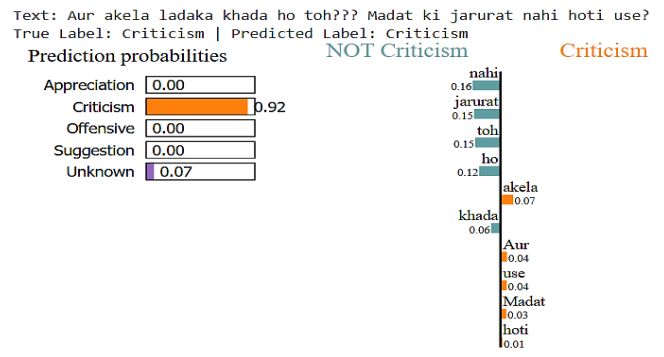

Fig. 10. Sample text for Criticism label

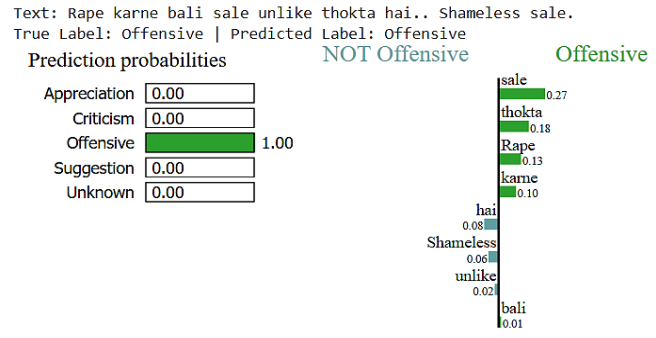

Fig. 11. Sample text for Offensive label

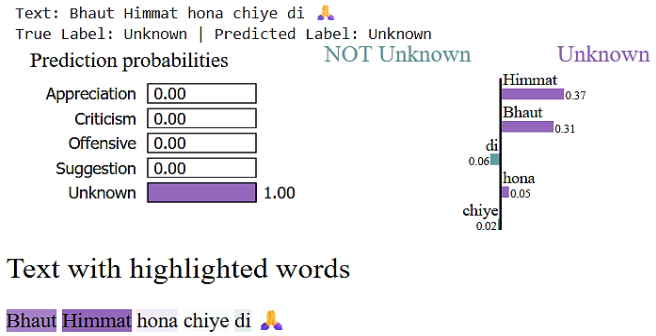

Fig. 12. Sample text for Unknown label

C. *Misclassified texts using the baselines and correctly classified by the proposed model to classify Subtask 1*

The translation of the example text “Har ladki ko education ka equal right milna chahiye, education se hi society aage badhegi.” is “Every girl should have an equal right to education; only through education will society progress.” It is actually classified as ‘optimistic’. The text was misclassified as ‘neutral’ using the FastText + BiGRU + Bahdanau attention model, but it was correctly classified as ‘optimistic’ by the proposed model. The misclassified and correctly classified cases are displayed in Figures 13 and 14, respectively.

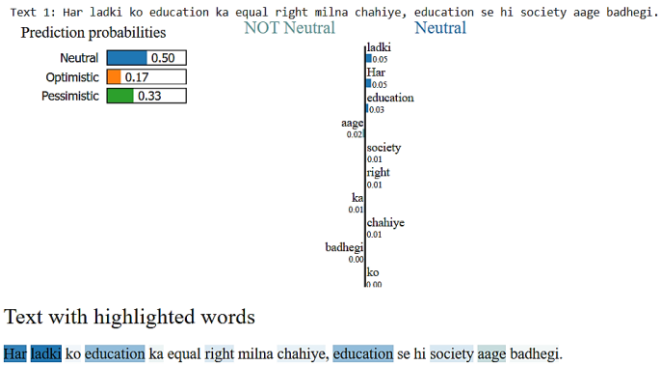

Fig. 13. Misclassified example text using the existing model

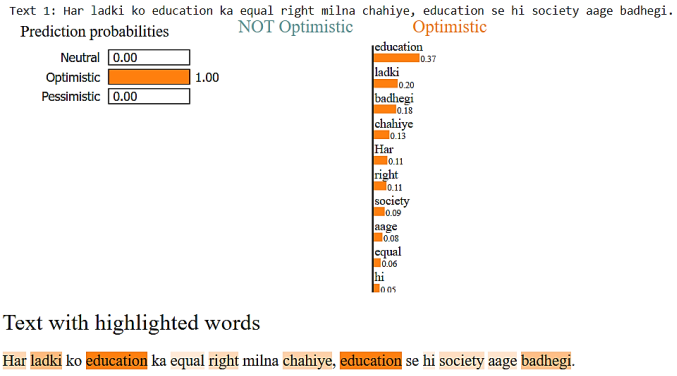

Fig. 14. Correctly classified example text using the proposed model

D. *Misclassified texts using the baselines and correctly classified by the proposed model to classify Subtask 2*

Another example text, “night mein bahar ghumne waale ladkiyon item hoti hai” states that the girls who roam outside at night are ‘items’. The word “item” in this context is a derogatory term used to objectify women. It is actually classified as ‘Criticism’. The text was misclassified as ‘Unknown’ using the FastText + BiGRU + Bahdanau attention model, but it was correctly classified as ‘Criticism’ by the proposed model. The misclassified and correctly classified cases are displayed in Figures 15 and 16, respectively.

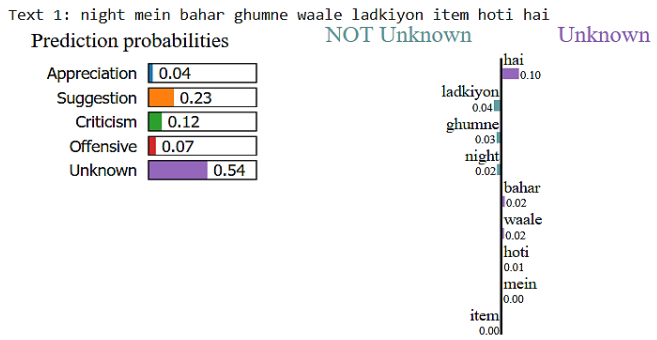

Fig. 15. Misclassified example text using the existing model

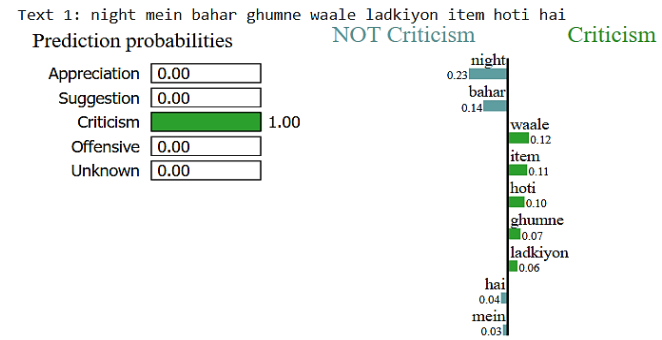

Fig. 16. Correctly classified example text using the proposed model

VI. Conclusion and Future Work

The experiments are conducted on a single benchmark dataset focused on misogyny detection in multilingual and code-mixed social media text. Byte-level ByT5 embeddings can handle inconsistent spellings and transliteration variations by encoding text independently of predefined vocabularies, making them robust to orthographic noise. The BiGRU component explicitly models bidirectional contextual dependencies. The Bahdanau attention mechanism enables the model to focus on the most salient parts of the input that impact the performance of misogyny classification. This hybrid architecture addresses noise robustness, sequential modeling, and interpretability, which overcome the limitations of the baseline models.

The integration of ByT5 embeddings with BiGRU and attention layers can increase training time and memory consumption, leading to computational overload. Although byte-level representations improve robustness to orthographic noise, they may overlook higher-level semantic regularities when trained on limited data, potentially affecting generalization in low-resource settings. These limitations suggest that further validation across large balanced datasets is needed to assess the scalability and generalization of the proposed approach.

In the future, multimodal data will be considered for misogyny classification capable of handling both text and image inputs by implementing vision transformers. The multi-head and hierarchical attention mechanisms are explored to capture subtle linguistic cues of misogyny across multilingual and code-mixed texts. A larger number of balanced datasets will be used to train the model, thereby avoiding data imbalance and yielding maximal accuracy and minimal loss. Large language models such as GPT, LLaMA, and Mistral models have to be explored for multilingual and code-mixed text using zero-shot and few-shot misogyny detection, compared with current hybrid approaches. Implementing such models can make social media more secure for women and save them from a misogynistic environment in real-world applications.

References

[1] K. I. Kumar, G. Sthanusubramoniani, D. Gupta, A. R. Nair, Y. A. Alotaibi, and M. Zakariah, "Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches," *Egyptian Informatics Journal*, vol. 30, p. 100683, Apr. 2025, doi: 10.1016/j.eij.2025.100683.

[2] A. Singh, D. Sharma, and V. K. Singh, "Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models," *Computer Speech & Language*, vol. 89, p. 101682, Jun. 2024, doi: 10.1016/j.csl.2024.101682.

[3] S. Saumya, A. Kumar, and J. P. Singh, "Filtering offensive language from multilingual social media contents: A deep learning approach," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108159, Feb. 2024, doi: 10.1016/j.engappai.2024.108159.

[4] M. Neog and N. Baruah, "A hybrid deep learning approach for Assamese toxic comment detection in social media," *Procedia Computer Science*, vol. 235, pp. 2297–2306, Jan. 2024, doi: 10.1016/j.procs.2024.04.218.

[5] N. Khanduja, N. Kumar, and A. Chauhan, "Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation," *Systems and Soft Computing*, vol. 6, p. 200112, Jun. 2024, doi: 10.1016/j.sasc.2024.200112.

[6] R. Rajalakshmi, S. Selvaraj, F. M. R, P. Vasudevan, and A. K. M, "HOTTEST: Hate and Offensive content identification in Tamil using Transformers and Enhanced STemming," *Computer Speech & Language*, vol. 78, p. 101464, Oct. 2022, doi: 10.1016/j.csl.2022.101464.

[7] P. K. Roy, S. Bhawal, and C. N. Subalalitha, "Hate speech and offensive language detection in Dravidian languages using a deep ensemble framework," *Computer Speech & Language*, vol. 75, p. 101386, Apr. 2022, doi: 10.1016/j.csl.2022.101386.

[8] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, "Hate speech detection on Twitter using transfer learning," *Computer Speech & Language*, vol. 74, p. 101365, Feb. 2022, doi: 10.1016/j.csl.2022.101365.

[9] B. R. Chakravarthi, M. B. Jagadeeshan, V. Palanikumar, and R. Priyadharshini, "Offensive language identification in dravidian languages using MPNet and CNN," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100151, Dec. 2022, doi: 10.1016/j.jjimei.2022.100151.

[10] A. Akhter, U. K. Acharjee, Md. A. Talukder, Md. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for Bengali cyber bullying detection in social media," *Natural Language Processing Journal*, vol. 4, p. 100027, Jul. 2023, doi: 10.1016/j.nlp.2023.100027.

[11] E. Hashmi, S. Y. Yayilgan, and S. Shaikh, "Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers," *Social Network Analysis and Mining*, vol. 14, no. 1, Apr. 2024, doi: 10.1007/s13278-024-01245-6.

[12] M. A. Rosid, D. O. Siahaan, and A. Saikhu, "Sarcasm detection in Indonesian-English Code-Mixed text using multihead Attention-Based Convolutional and Bi-Directional GRU," *IEEE Access*, vol. 12, pp. 137063–137079, Jan. 2024, doi: 10.1109/access.2024.3436107.

[13] S. Biradar, S. Saumya, and S. Kavatagi, "Safeguarding the Integrity of Online Social Networks (OSN): Leveraging the Efficacy of Conv-LSTM based Siamese Network to predict hate speech in low resource Hindi-English code-mixed text," *IEEE Access*, p. 1, Jan. 2025, doi: 10.1109/access.2025.3597144.

[14] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "ByT5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, 2022, doi: 10.1162/tacl_a_00462.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 1–16, 2014, doi: 10.1162/tacl_a_00461.

[16] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 577–585, arXiv:1506.07503.

S. Karishma received her M.Tech. Degree in Information Security from Puducherry Technological University, Puducherry, India. She is currently pursuing research at Puducherry Technological University, Puducherry, India. Her research interests include Information Security, Data Science, and Natural Language Processing.

V. Akila received her Ph.D. degree from Pondicherry University, Puducherry, India. She is currently working as an Associate Professor at Puducherry Technological University, Puducherry, India. She has authored and co-authored more than 50 publications. Her research interests include Social Network Analysis, Data Analytics, Spam and Botnet detection. She is currently working on an ongoing SERB-funded project titled 'Misinformation Black Propaganda'.