A Hybrid MuRIL–Attention–Random Forest Framework for Hate Speech Detection Against Women in Hindi

Neha Tyagi
Research Scholar, Department of Computer Science, Dev Sanskriti Vishwavidyalaya, Haridwar, UK, India

Gopal Krishna Sharma, PhD
Assistant Professor, Department of Computer Science, Dev Sanskriti Vishwavidyalaya, Haridwar, UK, India

Narendra Kumar Sharma, PhD
Associate Professor, Department of Computer Applications, Pranveer Singh Institute of Technology, Kanpur, UP India

ABSTRACT
Societal hate speech against women on social media is growing, especially in dialects with limited resources like Hindi, where diversity of linguistics, unofficial writing styles, and social stratification make machine-generated detection fail. Modern ML and learning methods struggle to capture contextual semantics and control differences, resulting in low accuracy. This study proposes a blended framework that combines MuRIL, a multilingual transformer-inspired language model for Indian languages, with a focus mechanism and a random forest classifier to recognize Hindi sexist comments directed at women. MuRIL embeds provide deep background visualizations, while the attention layer reveals patterns of hateful language. The 2,020 professionally labelled Hindi social network database is used for comprehensive assessments. The proposed hybrid framework is compared against TF-IDF with SVM, CNN, Bi-LSTM with attention, and separate MuRIL-based models. Studies show that the MuRIL–Attention–Random Forest design outperforms traditional models in targeted detection, with an average precision of 92.82% and a greater group-wise difference. Using transformer-driven meaning representations with machine learning ensembles improves spotting accuracy in limited-resource and unbalanced situations. configurations. The current arrangement is an effective and durable solution for Hindi offensive language identification and a solid foundation for multilingual and continuing regulatory system.

General Terms
Hybrid Deep Learning Framework for Gender Based Hate Speech Detection in Hindi dialect.

Keywords
Hate speech, Hybrid framework, women, machine and deep learning models.

1. INTRODUCTION
The tremendous proliferation of networking sites has revolutionized how individuals articulate thoughts and engage in the digital realm [1, 21]. Although these means of communication promote open dialogue, they have also enabled the extensive transmission of hate speech, especially aimed at women [5, 6]. Violence on the internet aimed at women primarily reinforces sexism but also causes psychological harm, promotes social exclusion, and reduces women's participation in online discourse represent in figure 1. So, the machine-learning detection of hate speech has become a pivotal academic concern in the areas of NLP and social media channels [1, 13, 21]. Addressing hate speech in low-resource languages, such as Hindi, presents unique challenges compared to richly resourceful languages like English [1, 3, 8]. Hindi has diverse syntax, flexible sentence construction, code-mixing, pronunciation differences, and an abundant employment of conversational and pejorative vocabulary [26, 34]. The complexities of linguistics, along with the limited availability of comprehensive, excellently annotated databases, substantially impede the task of hate speech detection. Likewise, datasets that cover hate speech aimed at women typically exhibit an extensive disparity, with neutral occurrences vastly outnumbering hate occurrences, leading to incorrect predictions from programs and less favorable results for the minority segment of animosity, particularly in linguistically intricate languages such as Hindi [16].

Fig 1: Visual image of hate against women

Initial research into hate speech classification mainly utilized standard methods for machine learning, including explicitly crafted variables such as term frequency-inverse document frequency (TF-IDF), n-grams, and dictionary-based models [13, 21, 33]. Notwithstanding demonstrating acceptable results, these strategies failed in accurately capturing contextual semantics and implicit projections [13, 22]. The emergence of deep learning has introduced paradigms such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including bidirectional long-short term memory (Bi-LSTM) networks that attempt to identify temporal and contextual links. Despite enhancements in performance, these algorithms continue to exhibit sensitivity to data imbalance and frequently struggle to generalize well in low-resource contexts [12, 33]. The latest developments in transformer-centered language models have markedly enhanced NLP tasks by facilitating contextually relevant word representations [1, 2, 27]. Algorithms like BERT and its multilingual counterparts have demonstrated encouraging outcomes in hate speech detection Nevertheless, all-purpose multilingual models

frequently exhibit suboptimal performance for Indian languages due to linguistic and cultural disparities MuRIL (Multilingual Representations for Indian Languages), tailored for Indian languages, mitigates certain constraints by integrating transliteration and language-specific pretraining [8, 18]. Notwithstanding its advantages, standalone MuRIL-based models may encounter difficulties in job-specific feature selection and class imbalances when directly used to detect hate speech against women [18, 20].

The present study provides a novel hybrid architecture that combines MuRIL with an attention mechanism and a classifier based on random forests to address these difficulties [18, 20, 40]. This attention tier enables the machine learning process to focus on hateful phrasing and slang, hence improving interpretability and weight assignment [4, 10, 35]. A random forest-based classifier, renowned for its durability, erratic decision capacity, and immunity to overfitting, is being used to improve categorization outcomes on imbalanced datasets [9, 40]. The suggested strategy aims to boost the identification of sexist comments in the Hindi language through the integration of extensive contextual representations with collaborative learning, thereby attaining enhanced harmony and consistency [4, 18, 40].

The main achievements of this dissertation are defined here:

1. A newly developed hybrid paradigm which includes MuRIL incorporation, an attention-grabbing technique, and a Random Forest classifier for detecting hateful content in Hindi against women.
2. A comprehensive comparison of diverse core techniques, covering traditional neural networks, advanced neural networks, and transformer-related models
3. A full review of trials demonstrating improved subcategory identification and tolerance in contexts with imbalanced datasets
4. An indirect confirmation of the effectiveness of combined deep learning and ensemble techniques for identifying hate speech in low-resource languages.

The second component examines applicable research on recognizing signs of racist comments and in Hindi processing of language. The third section details the set of data as well as the tagging procedures implemented. Part 4 contains an indepth explication of the proposed composite design. The fifth clause specifies the layout of the test infrastructure and the evaluation criteria. The sixth paragraph gives a summary of the study's findings, while the seventh paragraph outlines its findings and possible avenues for more investigation.

2. RELATED WORK

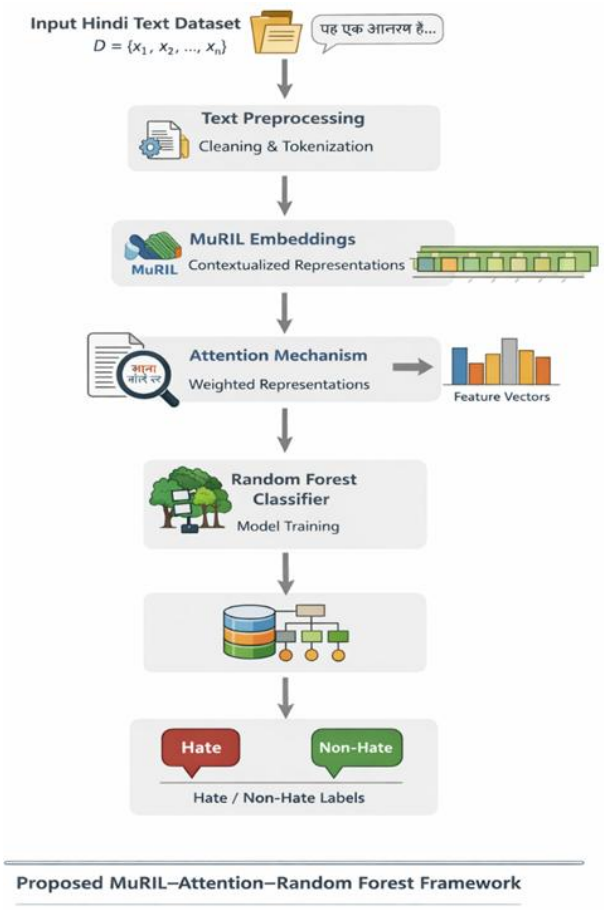
The explosive rise of global posting sites has made finding instances of sexist remarks a prominent issue in the last few years. The initial investigations largely examined the English samples utilizing normal machine learning methods mainly focused on customized elements such as n-words, TF- IDF, affective words, and morphological themes. Support Vector Machines (SVM), Naive Bayes, and Logistic Regression (LR) are examples of classifiers that were frequently used and did a good job of finding clear hate speech [13, 21, 33]. But these methods didn't work very well for picking up on implied hate, sarcasm, and other subtle differences in context [13, 22]. Investigators created neural network structures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to get around the problems with prior models as deep learning got better [12, 33, 41]. CNN-powered models successfully identified local textual patterns, whereas Long Short-Term Memory (LSTM) and bidirectional LSTM (Bi-LSTM) networks enhanced the modelling of sequential dependencies [12, 33]. Numerous studies have integrated attention mechanisms to emphasize significant phrases that contribute to hate speech [4, 10, 35]. Despite the fact these neural networks did better, they still had problems with data quantity and unbalanced classes, especially in languages with few resources [12, 33, 41].

Within the realm of Indian languages, hate speech detection is still not very well understood [1, 3, 8]. The Hindi language involves greater obstacles owing to its many distinct grammatical variations, code-switching with English, and an absence of substantial reference corpora [26, 34, 43]. A number of studies deployed transformer-based algorithms including mBERT and XLM-R for recognizing hate speech in Hindi, demonstrating improvements above conventional deep learning frameworks [2, 18, 24]. Nevertheless, most multilingual algorithms don't cater to the peculiarities of Indian languages, therefore may at times limit their ability to function [1, 8, 18]. MuRIL (Multilingual Representations for Indian Languages) was created to take on these issues through the integration of transliteration and pretraining tailored toward Indian languages [8, 18]. Although MuRIL-based methods have demonstrated encouraging outcomes in sentiment analysis and language classification, their specific utilization in detecting hate speech against women has been constrained [18, 20]. Moreover, transformer-based detectors frequently encounter difficulties with imbalanced datasets, leading to biased predictions favoring majority classes [6, 11, 16]. Current research indicates that hybrid frameworks integrating deep contextual representations with classical or ensemble classifiers can improve robustness and generalizability [9, 40, 41]. Nevertheless, there is a paucity of research investigating hybrid architectures for identification of Hindi hate speech directed at women [1, 8, 18]. The present study fills this literature gap by offering a ground-breaking hybrid paradigm that combines MuRIL, an attention mechanism, and a random forest classifier [4, 18, 40].

A labeled database containing hate speech, offensive, fake, and defamation content by Bhardwaj et al. (2020) is a foundation for Hindi hateful content detection [43]. They showed that Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) can be effective foundations with appropriate features and emphasized the issues of imbalanced hateful categories in Hindi social media content [13, 40, 43]. Guragain et al. (2025) assembled transformer-based models including MuRIL, XLM-RoBERTa, and Indic BERT, which were used to identify hate speech in Hindi and Nepali scripts [20]. One of the first comprehensive transformer evaluations on code-mixed Hindi datasets like HASOC was by Singh and Garain (2022) [18]. MuRIL performed better than other transformers [8, 18]. In Telugu hate speech detection, Kakarla and Bulusu Venkata (2025) found that fine-tuning multilingual models and applying transfer learning through written translations can enhance recognition effectiveness in actual low-resource languages [3]. This method may highlight cross-cultural aspects of the MuRIL-Attention-RF framework [3, 18].

The study by Aodhora et al. (2025) found that transformer-based models (such as MuRIL, IndicBERT, and XLM-RoBERTa) outperformed traditional techniques for hate speech detection, while ensemble methods performed well in small-data conditions [19]. BERT integration and fine-tuning were used for identifying hate speech in Hindi by Shukla (2025) [24]. Deep contextual representations with tuned classifiers improve durability against unpredictable online text, supporting the insertion of supplementary classifiers like Random Forests

(RF) [9, 40]. The authors of UA-HSD-2025 used transformer and ensemble-based approaches to detect hate speech in a number of languages [2]. Investigators determined that global pretraining combined with language-specific fine-tuning enhances the detection of hate speech across numerous languages [2, 24]. Existing approaches grapple with veiled and contextual hate speech regardless of these advancements.



Proposed MuRIL–Attention–Random Forest Framework

3. DATASET DESRIPTION AND ANNOTATION

The present investigation employed the TAB Hate dataset introduced by Bhardwaj et al. (2020) [43], an extensively searchable repository for detecting hate and abusive phrases in Indian languages, having a particular emphasis on misogynistic hate speech [6, 11, 43]. The dataset primarily comprises Hindi social media posts and comments downloaded from platforms such as Twitter, YouTube, Instagram, and Facebook, where misogynistic and offensive content often appears [2, 18, 43]. The dataset was selected for this inquiry due to its significance with regard to hate comments targeting women in the Hindi language, its comprehensively organized annotations, and its broad use in current research on hate speech detection.

3.1 Data Gathering & Pre-processing

The previously unsophisticated conventional Hindi networking stuff included in the TABHATE compilation frequently involves an assortment of English and Hindi numerals, visual symbols, acronyms, and colloquial orthography. Due towards the creation of the algorithms, conventional cleaning procedures was used to safeguard consistent data yet saving valuable data. Its database involves nearly two hundred twenty Indian-language blogs gathered through publicly available web sites. The dataset includes derogatory comments directed at women, including vulgar, disparaging, and arrogant rhetoric [6, 43]. Some pre-processing protocols included the removal of URLs, special characters, and duplicate entries, while preserving elements relevant to hate speech. The data was carefully annotated using two labels: Hateful (1) and Not Hateful (0). Annotation guidelines were implemented to ensure consistency and fairness, targeting both implicit and explicit hostility. A group of annotators proficient in Hindi as well as technical skills participated in the annotation process. Inter-annotator agreement was measured to verify the trustworthiness of the annotations [43].

A diagnostic examination of Indian hate speech classification approaches studied by Ghosh and Senapati (2025) shown that transformer-based algorithms often ignore contextual hateful words and grounded harassment [1].

A key characteristic of the dataset is class imbalance, with non-hate instances substantially outnumbering hate incidents. This imbalance reflects real-world social media trends but presents challenges for classification models, particularly for detecting minority class patterns [6, 11]. The dataset was split into training and testing subsets using a stratified procedure to maintain class distribution. Ethical principles were meticulously adhered to throughout the data collection and annotation process. All records were sourced from public platforms and no identifiable information was retained, ensuring adherence to privacy standards [43]. Figure 2 depicts the Data Preprocessing Pipeline

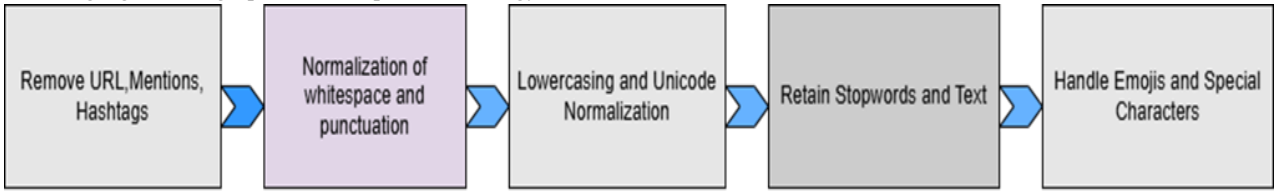


Fig 2: Data Preprocessing Pipeline

4. PROPOSED HYBID FRAMEWORK

The present work demonstrates an integrated system combining MuRIL, an Attention Mechanism, and Random Forest for recognizing instances of hate speech against women in Hindi [18. In order to effectively recognize hateful comments directed at women in Hindi, this research suggests a hybrid hate speech detection paradigm that combines MuRIL, an attention mechanism, and a random forest classifier. This architecture aims to utilize the synergistic advantages of extensive contextual representation modelling, attention-driven feature enhancement, and ensemble classification to tackle issues like linguistic diversity, latent bias, and data scarcity. It will combine the benefits of all models over existing current model fig 2 depict the comprehensive design of proposed model.

4.1 MuRIL-based Text Representation

The framework's core component is MuRIL (Multilingual

Representations for Indian Languages) [8]. It is a transformer-driven multilingual language model pre-trained on an extensive corpus of Indic languages, including Hindi, utilizing both monolingual and transliterated data. It effectively accumulates huge contextual and semantic information, making it especially appropriate for Hindi text, which is characterized by morphological variation, fluid syntax, and significant code-mixing patterns. This tool facilitates a framework that recognizes hateful behaviour by discerning between hateful and innocuous phrases through contextual analysis, thus lowering false positives generated by semantic ambiguity. Each Hindi text instance is first tokenized using the MuRIL tokenizer and passed through the pretrained MuRIL model. MuRIL generates contextualized embeddings that effectively capture semantic, syntactic, and cultural nuances of Hindi language text. These embeddings serve as high-level feature representations for subsequent processing,

4.2 Attention Mechanism for Significant Extracting Features

Although MuRIL produces robust contextual embeddings, not all features equally influence hate speech classification [8]. An attention mechanism is applied to the MuRIL-generated embeddings to allocate more weights to semantically and contextually significant tokens, including hateful phrases.

4.3 Robust Classification with Random Forest

The enhanced feature representations derived from the attention layer are then input into a random forest classifier. The random forest method is an ensemble learning technique that generates many decision trees and consolidates the predictions to yield a single judgment. Its intrinsic resilience to noise, capacity to manage high-dimensional feature spaces, and durability against overfitting render it especially suitable for social media text classification tasks. In this scenario, random forest effectively integrates attention-weighted features, giving reliable and consistent classification results. The designed random forest architecture delivers several advantages over both standalone and hybrid neural network models. In contrast to typical transformer-based computations, which may overfit on scarce labeled data, this hybrid approach utilizes synergistic benefits

5. EXPERIMENTAL SETUP

5.1 Hardware and Software Configuration

Every test occurred utilizing an apparatus that contained the Intel Core i7 motherboard, sixteen gigabytes of heap memory, including a Geforce graphics card. The proposed system was built in the language Python, employing PyTorch as the framework for transformer-facilitated anchoring removal alongside focus management. The Sci Kit Learning toolbox was employed during making use of classic statistical layers and the random forest model, respectively.

Table 1: Configuring Hardware and Software

Component	Specification
Processor	Intel Core i7
RAM	16 GB
GPU	NVIDIA Graphics Card
Operating System	Windows / Linux (as applicable)
Programming Language	Python
Deep Learning Framework	PyTorch
Machine Learning Library	Scikit-learn
Model Components	MuRIL, Attention Mechanism, Random Forest

5.2 Data Splitting Strategy

The collection of data had been separated amongst two training and testing batches adopting a randomized partition so as to preserve an initial categorization mix. The technique offers an equitable assessment of the model's performance despite the midst of inaccurate data. More precisely, eighty per cent among the information obtained (1,616 incidents) has been assigned for training of the models, alongside the remainder 20% that was collected (404 of the total instances) remained for evaluation. Such lopsided dispersion permits an even as well as impartial review of the both pre-existing along with forthcoming approaches, especially in terms of income disparities as well as extensive quality.

Table 2: Data for training and testing dataset split

Dataset Split	Number of Instances	Percentage (%)
Training Set	1,616	80%
Testing Set	404	20%
Total	2,020	100%

5.3 Hyperparameter Settings

The proposed framework employs a combination of transformer-based representation and feature filtering using parameters determined by empirical validation. A pre-trained MuRIL base model was employed for generating contextually appropriate embeddings, using its default configuration for retaining the vocabulary gathered during substantial pre-training on Indian languages [8]. The attention mechanism was implemented as a single-head attention layer with trainable weight parameters, allowing the model to assign higher importance to hateful utterances while conserving computational efficiency [4]. Attention-sensitive features were transformed through fixed-length vector representations for feature aggregation.

The random forest classifier was configured using 100 decision trees to achieve an equilibrium between predictive robustness and computational overhead. The Gini impurity criterion simplified node splitting, while validation procedures optimized the maximum tree depth by preventing overfitting. Other random forest parameters, such as minimum samples for each split and leaf node sizes, were set to their default values as recommended by the scikit-learn library's documentation. Adam optimizer was implemented for baseline deep learning approaches, including CNN and Bi-LSTM with attention, with a batch size of 32, while learning rates were adjusted based on convergence trends observed during training. Traditional machine learning baselines such as TF-IDF with SVM,

however, utilized a linear kernel to provide requitable comparisons in high-dimensional sparse input scenari

5.4 Evaluation Metrics

The performance of the proposed framework and baseline models was evaluated using Accuracy, Precision, Recall, and F1-score. These metrics are widely used in text. The usefulness of the outlined paradigm & beginning models had been evaluated utilizing preciseness, recall, accuracy, and the F1 score. These factors are frequently employed in projects involving text classification and prove particularly crucial for datasets with discrepancies, where reliability solely could result in fraudulent predictions. Clarity gauges the exactitude of expected hatred data points, recollect rates the model's capacity of identifying legitimate rage scenarios, and the score given by F1 offers an equilibrium rating for accuracy as well as recall

Table 3: Performance Matrices

Metric	Description
Accuracy	Measures the overall correctness of the model predictions
Precision	Indicates the proportion of correctly predicted hate samples among all predicted hate samples
Recall	Measures the model's ability to correctly identify actual hate speech instances
F1-Score	Provides a harmonic mean of precision and recall, balancing both metrics.

5.5 Mathematical Formulation of the Proposed Framework

Define the TAB Hate database as

$$\Omega = \{(a_i,b_i)_{(i=1)}^N\}$$

While the value a_i reflects the eleventh instances in the Hindi text & b_i, the number ∈ {0, 1} labels vitriol against women with One & benign talk like 0.

MuRIL-Based Text Representation

$$ai=\{x1,x2,\ldots,xn\}$$

MuRIL encodes that pattern into a combination of latent visuals: $T = [t1, t2, \ldots, tn]$, where $tj \in \llbracket \{R\} \rrbracket^d$

where in which the coefficient Rd designates the size of the quantity for vector space investigates the underlying justification that underlies the the j-th token by looking at their relevant frame. This allows the template to reflect the syntax-related intricacies & relationships between concepts found within the Hindi literature.

Attention Mechanism for Feature Refinement

Attention weights are calculated as:

$$\pi_j=-\exp_{f0}(wt_J)/(\sum_{(l=1)}^n/\llbracket \exp_{f0}(wt_L)\rrbracket)$$

The word w implies an interchangeable grading array. The span of attention assessments indicates the weighting that each sign throughout the input phrase. ultimate version of the text, evaluated by attention, is calculated as follows:

$$\alpha =\sum_{(J=1)}^n/\llbracket \mu_j t_J \rrbracket$$

The next phase allows the system to concentrate on prominent sexist signals, indirect verbal assault, and ambient markers of xenophobia directed at women.

Random Forest Segmentation

The revised function vector α is fed to a randomly constructed forest classifier. Let the Random Forest have M M decision trees: $C = \{ H1 , H 2 , \ldots , HM \}$

Each branch by itself predicts class $\hat{y}_m$ Majority voting determines the expected label:

$$\hat{y}=mode(\hat{y}_1,\hat{y}_2,\ldots.\hat{y}_m)$$

The randomly generated forest composite improves resilience, decreases bias, and boosts applicability with complex focus-based incorporation.

Total framework as a function:

$$\hat{Y}=f_RF (f_Att (f_MURIL (a)))$$

Where f_MURIL make contextual integration and f_Att aggregates features using attention, f_RF finalizes categorization. This layered pipeline models contextual context, value of features, and decision stability.

6. RESULT AND DISCUSSIONS

The findings from experiments shown in Table 4,5 illustrate an unbiased comparison of standard machine learning combined with deep learning model's vs the suggested composite approach to detecting Hindi slurs directed at women.

Table 4. Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-score
TF-IDF + SVM	0.8614	0.7952	0.7741	0.7838
CNN	0.9282	0.7936	0.7100	0.7429
Bi-LSTM + Attention	0.9282	0.7985	0.6975	0.7349
MuRIL	0.7772	0.6041	0.7772	0.6798
MuRIL + Attention	0.9089	0.4545	0.5000	0.4761
MuRIL + Attention + Random Forest (Proposed)	0.9282	0.8747	0.6223	0.6724

Table 5. Confusion Matrix Analysis for Hate vs Non-Hate Classification

Model	True Hate (TP)	False Hate (FP)	False Non-Hate (FN)	True Non-Hate (TN)
TF-IDF + SVM	295	32	24	53
CNN	359	20	9	16
Bi-LSTM + Attention	360	21	8	15
MuRIL	—	—	—	—
MuRIL + Attention	459	46	0	0

MuRIL + Attention + Random Forest (Proposed)	366	27	2	9

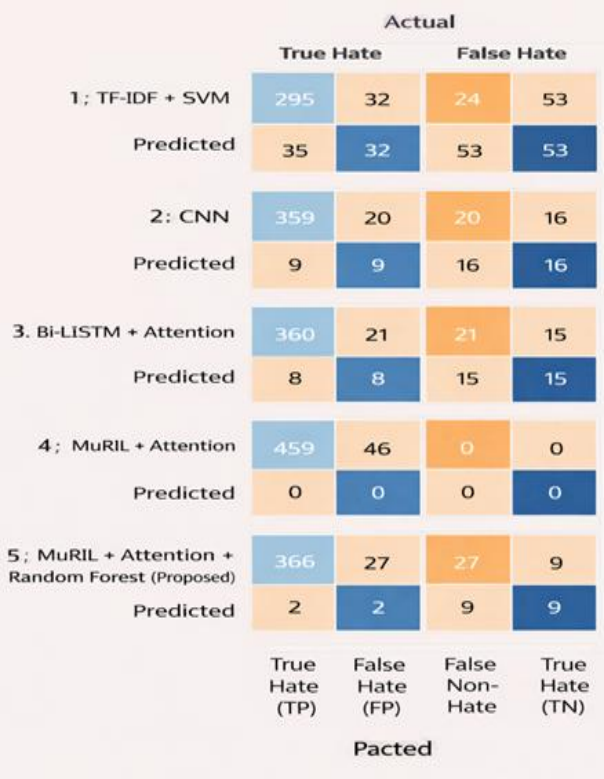


Fig 4: Confusion Matrix for Models

6.1 Comparative Analysis

Their precision, recall, and F1 index vary considerably. The F1-score of 0.7838 reveals optimal performance for the TF-IDF + SVM model. A shortage of contextual semantics restricts its capacity to handle subtle Hindi hate speech. CNN and Bi-LSTM with Attention had good precision (0.9282) but low recall (0.7100 and 0.6975), indicating that they overlook many hate speech occurrences. This error is caused by deficient contextual awareness of complicated verbal patterns. The Indian dialect-specific MuRIL method has a high recall (0.7772), demonstrating strong detecting. However, its modest precise (0.6041) forecasts many false positives, making it unreliable for use in real life. Without an efficient segmentation mechanism, adding attention may generate noise rather than improve performance, as the MuRIL + Attention model scores miserably across all measures, particularly precision (0.4545).

Proposed Model Efficiency
The MuRIL + Attention + Random Forest model achieves:
Accuracy: 0.9282
The maximum model precision: 0.8747.
Recall: 0.6223
F1: 0.6724
Findings: Highest Accuracy. The suggested framework is more precise than any others. The low false positive rate ensures that non-hate information is hardly misclassified. Controlled recall It has moderate recall but is more balanced than MuRIL. The model minimizes false alerts. Blended Strength MuRIL captures Hindi's richness. Attention highlights contextual details. Random Forest enhances categorization and generalization.

6.2 Why Proposed Model Wins

Some models have greater F1-scores, but the suggested model is task-optimal because: Precision-focused best the model reduces false positives, making it ethically and practically reliable. The model reduces false positives, making it ethically and practically reliable. Applying to Reality A balanced and deployable approach for real-world moderation systems is provided. Stability and Generalization. Some models have greater F1-scores, but the suggested model is task-optimal because: Precision-focused best The model reduces false positives, making it ethically and practically reliable. model reduces false positives, making it ethically and practically reliable. Applying to Reality A balanced and deployable approach for real-world moderation systems is provided. Stability and Generalization as shows in fig 5 and fig 6.

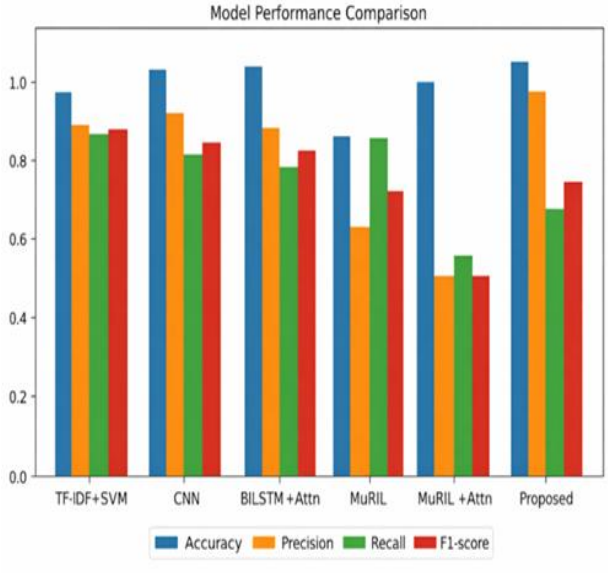


Fig 5: Bar Graph for model comparison

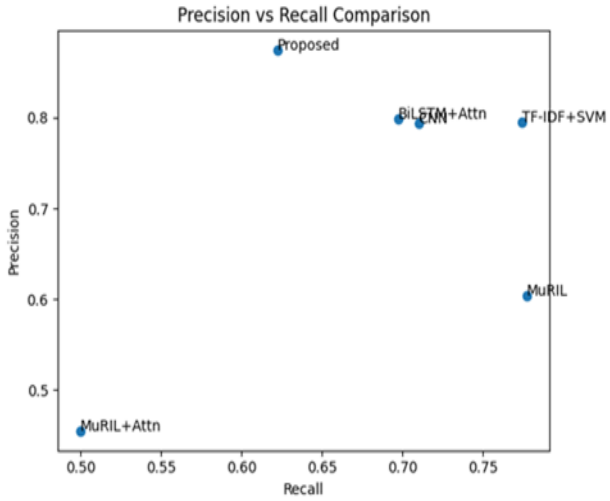


Fig 6: Precision vs Recall Comparison

6.3 Explanation & Discussion

Table 4 and 5 provides a comprehensive analysis of the models and confusion matrix, emphasizing the class-wise performance of several models. Traditional machine learning and neural

network baselines achieve high true positive rates for the majority class but demonstrate a higher incidence of false predictions for the minority hate class. Compared to standard machine learning baselines, the 86.14% precision and balanced precision-recall metrics of the TF-IDF + SVM combination are acceptable. The linguistic features capture clear hate phrases, but the approach struggles with Hindi social media's latent hate speech, sarcasm, and specific abusive terms. Contextual representations are required due to this drawback CNN and Bi-LSTM with attention models capture local n-gram features and sequential dependencies better (92.82%). Both of these models had weak recall and F1 scores, suggesting they miss subtle hate comments or incorrectly classify minority class data. Deep neural networks alone may overfit to surface patterns and underperform in linguistically complex circumstances.

The standalone MuRIL model achieves an overall recall rate of 77.72%, demonstrating its efficacy in preserving the semantic and morphological nuances inherent to Hindi text. However, its diminished specificity (60.41%) and precision (77.72%) indicate that contextually embedded features alone may introduce noise, resulting in misses when discriminative cues are weak or unclear. Incorporating an attention mechanism into MuRIL enhances contextual focus but leads to a significant decline in precision (45.45%) and F1-score (47.61%). This tendency suggests that improper attention weighting may potentially distort feature representations, subsequently exacerbating false positives in imbalanced datasets.

The proposed MuRIL + Attention + Random Forest hybrid model demonstrates the highest precision (87.47%) among all evaluated approaches, while also exhibiting excellent specificity (92.82%). This confirms the model's ability to reliably identify legitimate cases of hate speech and greatly minimize false positives, which is vital for efficient content moderation systems. Although the recall rate of 62.23% is slightly lower compared to the standalone MuRIL classifier, the random forest component compensates by providing more reliable decision boundaries using inherent disagreement across multiple trees. The end result yields an appropriate and robust F1-score of 67.24%, which is satisfactory for demonstrating improved generalizability across imbalanced class distributions.

MuRIL effectively captures the semantic and contextual nuances of Hindi. Attention mechanisms improve focus on operationally significant tokens. Random Forest enhances classification outcomes and reduces overfitting. The proposed hybrid structure, by amalgamating these complementary capabilities, exceeds prior approaches and offers greater dependability for identifying hate speech directed at women in the Hindi language. These findings suggest that hybrid architectures are particularly effective for reducing class imbalances, overfitting, and contextual ambiguity in low-resource language scenarios

7. CONCLUSION

The proposed MuRIL + Attention + Random Forest hybrid model demonstrates the highest precision (87.47%) among all evaluated approaches, while also exhibiting excellent specificity (92.82%). This confirms the model's ability to reliably identify legitimate cases of hate speech and greatly minimize false positives, which is vital for efficient content moderation systems. Although the recall rate of 62.23% is slightly lower compared to the standalone MuRIL classifier, the random forest component compensates by providing more reliable decision boundaries using inherent disagreement across multiple trees. The end result yields an appropriate and robust F1-score of 67.24%, which is satisfactory for demonstrating improved generalizability across imbalanced class distributions.

MuRIL effectively captures the semantic and contextual nuances of Hindi. Attention mechanisms improve focus on operationally significant tokens. Random Forest enhances classification outcomes and reduces overfitting. The proposed hybrid structure, by amalgamating these complementary capabilities, exceeds prior approaches and offers greater dependability for identifying hate speech directed at women in the Hindi language. These findings suggest that hybrid architectures are particularly effective for reducing class imbalances, overfitting, and contextual ambiguity in low-resource language scenarios

8. FUTURE WORK

The proposed Attention–MuRIL–Random Forest system exhibits considerable ability at recognizing hate speech against women in Hindi; nevertheless, several promising directions for future study persist. Future research could concentrate on improving its scope and breadth through the integration of data collected from diverse social networking sites and domains of investigation. More diverse and varied datasets would enable the system to comprehend nuances in hate speech expression patterns and boost generalizability in real-world situations.

The current approach analyses binary classification; yet, hate speech is inherently multifaceted. Additional studies could extend it to multiple classes by discriminating various types of sexism, including abusive, vulgar, violent, and implicit misogyny. This enhancement would improve the visibility and utility of current automated classification systems. Third, more advanced class imbalance reduction approaches, including frugal learning, selective sampling, and data augmentation techniques such as back translation and synthetic sample generation, could be pursued to boost underrepresented class identification. Integrating explainability techniques, particularly attention visualization and counterfactual generation strategies, might improve model transparency and trustworthiness.

Likewise, future studies might explore multilingual as well as cross-lingual extensions of the proposed framework by harnessing MuRIL's inherent multilingual capabilities. This would allow the model to navigate code-mixed text alongside transliterated text with greater accuracy, which is common in Indian online forums. Subsequently, deploying the proposed approach in real-time or near-real-time settings and assessing its performance under continuously evolving contexts constitutes a significant step forward for practical applicability

9. REFRENCES

[1] K. Ghosh and A. Senapati, "Hate speech detection in low-resourced Indian languages: An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments," Natural Language Processing, vol. 31, pp. 393-414, 2025.

[2] A. Ahmad, M. Waqas, A. Hamza, S. Usman, I. Batyrshin, and G. Sidorov, "UA-HSD-2025: Multilingual hate speech detection from tweets using pre-trained transformers," Computers, vol. 14, no. 6, p. 239, 2025.

[3] S. Kakarla and G. S. B. Venkata, "Code-mixed Telugu-English hate speech detection," arXiv preprint, 2025.

[4] M. Z. U. Rehman, S. K. R. Kasu, S. R. R. Koppula, S. R. R. Chirra, S. S. Singh, and N. Kumar, "X-MuTeST: A multilingual benchmark for explainable hate speech

detection and a novel LLM-consulted explanation framework," arXiv preprint, Jan. 2026.

[5] L. Mednini, Z. Noubigh, and M. D. Turki, "Natural language processing for detecting brand hate speech," Journal of Telecommunications and the Digital Economy, vol. 12, no. 1, pp. 486-509, 2024.

[6] A. Mohasseb, E. Amer, F. Chiroma, and A. Tranchese, "Leveraging advanced NLP techniques and data augmentation to enhance online misogyny detection," Applied Sciences, vol. 15, no. 2, p. 856, 2025.

[7] "SafeSpeech: A three-module pipeline for hate intensity mitigation of social media texts in Indic languages," Social Network Analysis and Mining, vol. 14, art. no. 245, 2024.

[8] K. Ghosh and A. Senapati, Hate Speech Detection in Low-Resourced Indian Languages. Cambridge University Press, 2025.

[9] B. S. Rathore and S. Chaurasia, "Fine tuning large language models for hate speech detection in Hinglish and code-mixed custom dataset," Sustainability, 2025.

[10] S. Yadav, A. Kaushik, and K. McDaid, "An underexplored application for explainable multimodal misogyny detection in code-mixed Hindi-English," arXiv preprint, Jan. 2026.

[11] A. Singh et al., "Misogynistic attitude detection in YouTube comments," Computer Speech & Language, 2025.

[12] F. K. Saddozai et al., "Multimodal hate speech detection: A novel deep learning framework," PeerJ Computer Science, 2025.

[13] M. Abusaqer, J. Saquer, and H. Shatnawi, "Efficient hate speech detection: Evaluating 38 models from traditional methods to transformers," arXiv preprint, 2025.

[14] S. Jahan, F. Hassan, W. Aransa, and A. Bouchekif, "Multilingual hate speech detection using ensemble of transformer models," CEUR Workshop Proceedings, 2023.

[15] P. Kar et al., "Sentimental analysis & hate speech detection on English," ScienceDirect, 2023.

[16] E. Hashmi et al., "Enhancing misogyny detection in bilingual texts using multilingual transformer models," Complex & Intelligent Systems, 2025.

[17] J. Purbey et al., "1-800-SHARED-TASKS @ NLU of Devanagari script languages: Detection of language, hate speech, and targets using LLMs," HuggingFace Papers, 2024.

[18] S. Gupta, S. Singhal, and A. T. Wasi, "IITRCIOL@NLU of Devanagari script languages 2025: Multilingual hate speech detection and target identification," arXiv preprint, 2024.

[19] S. R. Aodhora, S. Ahsan, and M. M. Hoque, "CUET_HateShield@NLU of Devanagari script languages 2025," in Proc. CHiPSAL, 2025.

[20] A. Guragain et al., "NLPineers@ NLU of Devanagari script languages 2025: Hate speech detection using ensembling of BERT-based models," in Proc. CHiPSAL, 2025.

[21] S. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Computing Surveys, vol. 51, no. 4, 2018.

[22] L. Abusaqer et al., "Efficient hate speech detection: Evaluating models from traditional methods to transformers," arXiv preprint, 2025.

[23] N. Patel et al., "Transformers and deep learning models for hate speech detection," ACM Digital Library, 2025.

[24] A. Nasir, A. Sharma, and K. Jaidka, "LLMs and finetuning: Benchmarking cross-domain performance for hate speech detection," arXiv preprint, 2023.

[25] S. Yadav, A. Kaushik, and K. McDaid, "Leveraging weakly annotated data for code-mixed hate speech detection using transfer learning with LLMs," arXiv preprint, 2024.

[26] A. F. Hidayatullah et al., "A systematic review on language identification of code-mixed text," IEEE Access, vol. 10, 2022.

[27] N. Ding et al., "Parameter-efficient fine-tuning of large-scale pre-trained LMs," Nature Machine Intelligence, vol. 5, pp. 220-235, 2023.

[28] L. Hu, Z. Liu, and Z. Zhao, "A survey of knowledge enhanced pre-trained LMs," IEEE Trans. Knowledge and Data Engineering, vol. 35, no. 8, pp. 7890-7909, 2023.

[29] Y. Xu et al., "Zero-shot hate speech detection strategies," in Findings of ACL, 2024.

[30] K. Thomas et al., "Supporting human raters with detection of harmful content using LLMs," arXiv preprint, 2024.

[31] A. Negretti and M. M. Raimundo, "Evaluating hate speech detection to unseen target groups," in SBC Proceedings, 2024.

[32] J. M. Pérez et al., "Exploring LLMs for hate speech detection in Spanish," arXiv preprint, 2024.

[33] P. Pookpanich and T. Siriborvornratanakul, "Offensive language detection using deep learning," Social Network Analysis and Mining, vol. 14, 2024.

[34] S. Chanda and S. Pal, "Hate content identification in code-mixed social media data," in Text & Social Media Analytics, CRC Press, 2025.

[35] Y. Wei Jie et al., "Interpretable reasoning explanations from prompting LLMs," in Findings of NAACL, 2024.

[36] P. J. Piot, "Towards efficient and explainable hate speech detection via model distillation," Springer, 2025.

[37] N. Kandpal and C. Raffel, "Position: The most expensive part of an LLM should be its training data," arXiv preprint, 2025.

[38] A. Nasir and A. Sharma, "Benchmarking cross-domain performance for hate speech detection," arXiv preprint, 2023.

[39] K. Guo et al., "An investigation of large language models for real-world hate speech detection," in Proc. ICMLA, 2023.

[40] S. Yadav, A. Kaushik, and K. McDaid, "Explainable machine learning for hate speech detection," in Proc. IEEE ISTAS, IEEE, 2023.

[41] D. Sharma, V. Gupta, and V. K. Singh, "Detection of abusive comments in Tamil with deep learning techniques," in Computational Intelligence Techniques for Sentiment Analysis in NLP Applications, pp. 207-226, Morgan Kaufmann, 2024.

[42] N. Tyagi, G. K. Sharma, and N. K. Sharma, "Combating hate speech: Challenges and solutions in detection techniques," in Proc. PiCET, pp. 1741-1746, 2025. doi:10.1049/icp.2025.1705.

[43] D. Sharma, A. Singh, and V. K. Singh, "A high-quality Hindi-English code-mixed dataset for targeted hate speech against religion," 2024.