



# Context-Aware Hindi Text Classification for Digital Harassment Detection Using BERT and Traditional Models

<sup>1</sup>Jayshri Patel

<sup>1</sup>Assistant Professor,

<sup>1</sup>Department of Computer Science,

<sup>1</sup>Veer Narmad South Gujarat University, Surat, India.

**Abstract :** The growing prevalence of online hate speech and digital harassment in regional languages, particularly Hindi, necessitates the development of robust and context-aware detection systems. This study focuses on the Hindi subset of the HASOC dataset to design a binary classification framework for detecting digital harassment. We evaluate and access the performance of traditional machine learning models—Support Vector Machine, Random Forest, and Logistic Regression—with BERT, a transformer-based deep learning model. While traditional models rely on sparse features and statistical assumptions, BERT utilizes contextual embeddings and bidirectional attention mechanisms to capture the linguistic nuances of Hindi text. Experimental results show that BERT significantly outperforms the traditional classifiers in terms of accuracy and F1-score, highlighting its effectiveness for real-world applications such as content moderation and online abuse detection. These findings underscore the critical role of transformer-based models in advancing regional language processing and enhancing digital safety.

**Index Terms** - Digital harassment detection, Hindi hate speech, HASOC dataset, BERT, SVM, Random Forest, Logistic Regression, Context-aware classification, Offensive language, Cyberbullying.

## I. INTRODUCTION

Digital platforms have become central to modern communication, yet they are increasingly misused to spread hate, abuse, and offensive content. This trend is particularly alarming in multilingual societies like India, where regional languages—including Hindi—dominate user-generated content. Several studies, including those by Joshi et al. (2021) and Bhardwaj et al. (2020), have laid foundational work in hostility detection using deep learning in Hindi text. The detection of digital harassment in Hindi text presents unique challenges, such as informal expressions, code-mixing, and the lack of large-scale annotated datasets.

To address these issues, the HASOC (Hate Speech and Offensive Content) dataset has emerged as a benchmark for hate speech and offensive content detection in Hindi and other Indian languages (Basile, 2019). Specifically, the Hindi subset of the HASOC dataset provides well-annotated binary labels—HOF (Hate and Offensive) and NOT (Non-Offensive)—suitable for training and evaluating classification models in the context of digital harassment. Previous work by (Velankar, 2021) and (Bashar, 2020) has highlighted the importance of such multi-model evaluations for Indic languages. Moreover, recent contributions by (Patel, 2025) emphasize the role of AI in shaping proactive digital harassment detection frameworks.

Traditional machine learning models such as Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR) have been widely used for text classification due to their simplicity and interpretability. However, these models often fall short in capturing the contextual semantics required for detecting nuanced hate speech, especially in informal or code-mixed Hindi text.

In contrast, transformer-based models like BERT have shown state-of-the-art performance in multiple natural language processing tasks. With its bidirectional encoder structure and contextual embedding, BERT offers significant improvements in understanding semantic meaning and intent, making it ideal for detecting subtle forms of harassment in social media content.

In this study, we conduct a detailed performance assessment of traditional classification algorithms alongside the BERT deep learning model, utilizing the Hindi-only subset of the HASOC dataset—an established benchmark for hostile content detection in Indian languages. The primary goal is to explore how these models handle the linguistic complexity, code-mixing, and contextual nuances inherent in Hindi social media text. By integrating the interpretability and efficiency of traditional classifiers with the contextual depth offered by BERT's transformer-based architecture, the study seeks to improve the reliability, accuracy, and adaptability of automated content moderation systems. Ultimately, this research contributes to the development of more robust and scalable solutions for identifying digital harassment, thereby fostering safer and more inclusive online environments for Hindi-speaking communities.

## II. RELATED WORK

The automatic detection of hate speech and offensive language in Indian languages, particularly Hindi, has become a vital area of research due to the exponential rise of user-generated content on social media. Multiple studies have explored both traditional and deep learning techniques for this task, focusing on linguistic complexity, dataset challenges, and model performance.

An empirical evaluation of deep learning models for hostility detection in Hindi was conducted, highlighting the importance of dataset-specific tuning and demonstrating the effectiveness of transformer models when applied to the HASOC dataset (Joshi, 2021). A comprehensive hostility detection dataset in Hindi, which laid the foundation for robust model training and evaluation in regional languages, was developed and released (Bhardwaj, 2020). The use of convolutional neural networks (CNNs) for Hindi hate speech classification revealed that even simple deep learning models can outperform traditional approaches (Bashar, 2020).

A unified model designed to detect hate speech across multiple Indic languages was proposed, leveraging shared embeddings and cross-lingual transfer learning techniques (Bhatia, 2021). A focus on Indo-Aryan languages emphasized the challenges posed by dialectal diversity in user-generated content on social media platforms (Narayan, 2023). A comparative analysis of deep learning architectures for hate speech detection in Hindi and Marathi found transformer-based models to be superior in terms of contextual understanding (Velankar, 2021).

Globally, hybrid and multichannel deep learning strategies have been proposed for cyberbullying detection, with promising potential for regional adaptation (Alotaibi, 2021; Aldhyani, 2022). Efforts such as SemEval-2019 have provided essential datasets and multilingual frameworks for hate speech classification across languages (Basile, 2019).

Applied approaches to machine learning and AI strategies for cyberbullying and harassment detection in the Indian context were demonstrated in recent works (Deepana, 2023; Patel, 2025b). Additionally, practical insights into system deployment for digital harassment prevention using AI models were explored (Patel, 2025a).

Together, these studies underscore the need for a context-aware, scalable, and language-sensitive approach to digital harassment detection. Our work builds directly on these foundations by focusing exclusively on the Hindi HASOC dataset and comparing traditional classifiers with BERT.

## III. RESEARCH METHODOLOGY AND MODEL IMPLEMENTATION

A well-organized, multi-phase computational framework was systematically adopted in this study to effectively detect instances of digital harassment within Hindi-language social media content. The framework encompassed data preprocessing, feature engineering, model training, and evaluation, ensuring a comprehensive approach to the problem. The central objective of this research was to perform a comparative assessment between conventional machine learning algorithms and a state-of-the-art deep learning architecture—Bidirectional Encoder Representations from Transformers (BERT). Specifically, the study leveraged a benchmark dataset tailored for Hindi text classification, ensuring linguistic relevance and contextual richness in evaluating model performance across hostile and non-hostile content categories. By contrasting traditional models such as SVM and Random Forest with the contextual understanding capability of BERT, the research aimed to determine the most effective approach for identifying harassment in Hindi social media discourse.

### 3.1 Dataset Selection and Description

The HASOC 2021 Hindi dataset was selected as the foundational corpus for this research. It contains annotated social media posts labeled into two binary categories: HOF (Hate and Offensive) and NOT (Not Offensive). The dataset was filtered to retain only the necessary columns (text and task\_1) and cleaned by removing rows with missing values. The labels were converted to binary format (HOF as 1 and NOT as 0) for compatibility with binary classification algorithms.

### 3.2 Data Preprocessing

Standard Natural Language Processing (NLP) techniques were applied to enhance the quality and consistency of the textual data. The preprocessing steps included text normalization, wherein all characters were converted to lowercase to reduce redundancy, and removal of punctuation and special characters to eliminate noise. Additionally, commonly occurring stop-words in Hindi were filtered out to focus the model on informative terms. Tokenization was performed to prepare the text for model input by segmenting it into linguistically meaningful units. Finally, the target labels were encoded into binary numeric format, with offensive content (HOF) mapped to 1 and non-offensive content (NOT) mapped to 0, to facilitate supervised classification.

For the BERT-based deep learning model, tokenization was carried out using the pre-trained multilingual BERT tokenizer (bert-base-multilingual-cased), which employs sub-word segmentation and is designed to handle diverse linguistic structures, including those found in Hindi. This tokenizer allows the model to effectively capture syntactic and semantic nuances within the text, particularly in low-resource and multilingual contexts.

### 3.3 Feature Engineering

Two distinct approaches were used for feature representation:

**TF-IDF Vectorization:** Traditional models such as SVM, Random Forest, and Logistic Regression were trained using term frequency-inverse document frequency (TF-IDF) vectors. The maximum number of features was limited to 5000, and bigrams (n-gram range of 1–2) were included to capture contextual patterns.

**Contextual Embeddings (BERT):** For the deep learning model, BERT-generated contextual embeddings were used, leveraging the semantic relationships in text and capturing complex syntactic dependencies in the Hindi language.

### 3.4 Model Implementation

The machine learning models implemented for this research include both traditional and deep learning-based approaches. Implementation of SVM was carried out via the LinearSVC class from scikit-learn. The Random Forest Classifier was configured with 100 decision trees ( $n\_estimators = 100$ ), while Logistic Regression was trained using L2 regularization and set to a maximum of 1000 iterations to ensure convergence. In the deep learning setup, the BERT model was fine-tuned using the Hugging Face Transformers library, leveraging the BertForSequenceClassification class initialized with the bert-base-multilingual-cased pre trained weights.

All models were trained using an 80/20 train-test split to ensure unbiased evaluation and generalization. Traditional machine learning models underwent hyperparameter tuning through grid search combined with cross-validation to identify optimal configurations. BERT was trained using AdamW optimization with a 2e-5 learning rate for 3–5 epochs, and model selection was based on validation loss and F1-score.

### 3.5 Evaluation Metrics

To measure the performance of each classification model, several well-established metrics were employed. The confusion matrix components—including True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)—were calculated to evaluate the pattern of true and false predictions across the respective classes. In addition, quantitative performance measures such as accuracy, precision, recall, and F1-score were computed to assess the overall quality of each model's predictions.

These metrics offer a comprehensive perspective on the models' ability to accurately identify both offensive (HOF) and non-offensive (NOT) content, while managing the trade-off between false positives and false negatives. Special emphasis was placed on the F1-score, as it provides a balanced evaluation of precision and recall, and is particularly informative in scenarios with imbalanced class distributions, which are common in digital harassment detection tasks.

### 3.6 Model Development Workflow

Figure 1 presents a schematic overview of the proposed system workflow, encompassing all critical phases including dataset acquisition, preprocessing, feature engineering, model training, and performance evaluation. The diagram provides a structured visualization of the complete research pipeline and emphasizes the integration of both traditional machine learning techniques and transformer-based deep learning models (specifically BERT) for the task of Hindi-language digital harassment detection.

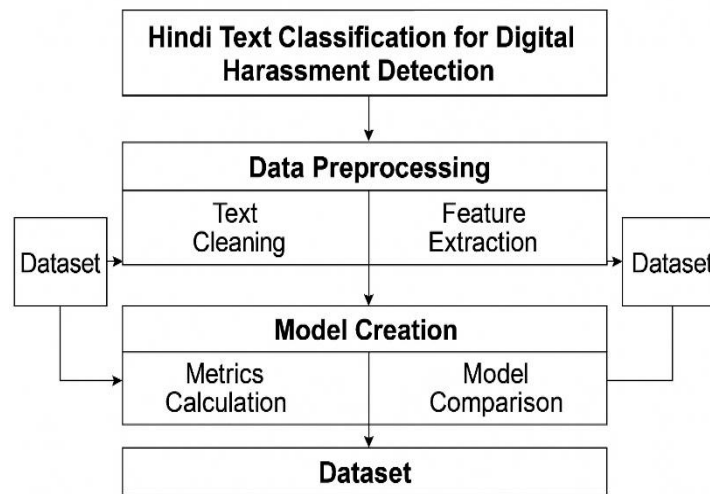


Figure 1: Context aware Hindi text Classification Model

## IV. RESULTS AND PERFORMANCE EVALUATION

This section presents the performance outcomes of four classification models—BERT, Support Vector Machine, Random Forest, and Logistic Regression—applied to the task of Hindi digital harassment detection using the HASOC dataset. Evaluation was performed using both confusion matrix components (True Positives, True Negatives, False Positives, and False Negatives) and standard performance metrics including accuracy, precision, recall, and F1-score.

The goal is to identify the best-performing model for offensive content detection, minimizing misclassifications. Results are summarized in Table 1.

Table 1: Evaluation of Classification Models for Hindi Digital Harassment Detection

| Model               | TP  | TN  | FP  | FN  | Accuracy | Precision | Recall | F1-Score |
|---------------------|-----|-----|-----|-----|----------|-----------|--------|----------|
| BERT                | 409 | 353 | 88  | 83  | 0.817    | 0.823     | 0.831  | 0.827    |
| SVM                 | 389 | 321 | 120 | 103 | 0.761    | 0.764     | 0.791  | 0.777    |
| Random Forest       | 403 | 278 | 163 | 89  | 0.730    | 0.712     | 0.819  | 0.762    |
| Logistic Regression | 384 | 326 | 115 | 108 | 0.761    | 0.770     | 0.780  | 0.775    |

The BERT model outperformed the other models across all metrics, achieving the highest accuracy (81.7%) and F1-score (0.827), with the fewest false positives (88) and false negatives (83). This suggests strong classification capability and generalizability, especially in real-world digital moderation scenarios. Both SVM and Logistic Regression recorded similar accuracy (~76.1%), though Logistic Regression achieved a slightly better balance between precision and recall. Random Forest showed high recall (81.9%) but had the lowest precision (71.2%) and the highest false positive rate, indicating a tendency to over-predict offensive content.

To visually interpret the classification distribution, Figure 2 presents pie charts representing the confusion matrix components (TP, TN, FP, FN) for each model, enabling quick assessment of each model's performance across correct and incorrect classifications.

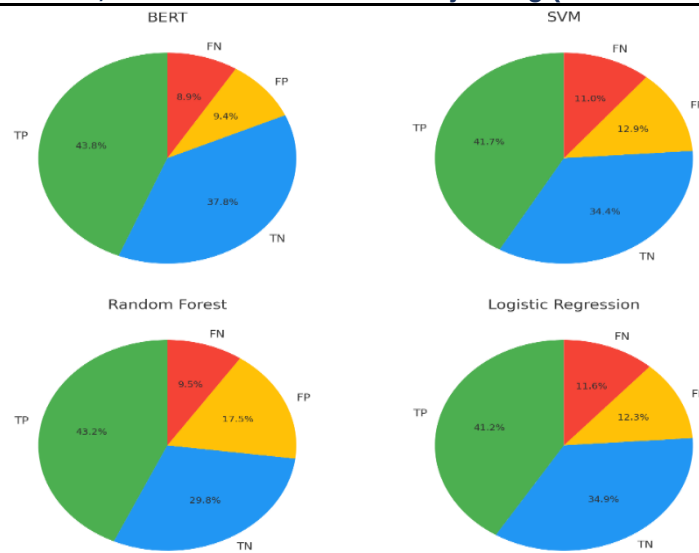


Figure 2. Confusion Matrix Distribution for All Models

In addition, a comparative bar chart is provided in Figure 3 to display the core performance metrics—accuracy, precision, recall, and F1-score—for all models in a unified format. This visualization highlights BERT's overall superiority in all categories, while also revealing the trade-offs between recall and precision in traditional classifiers.

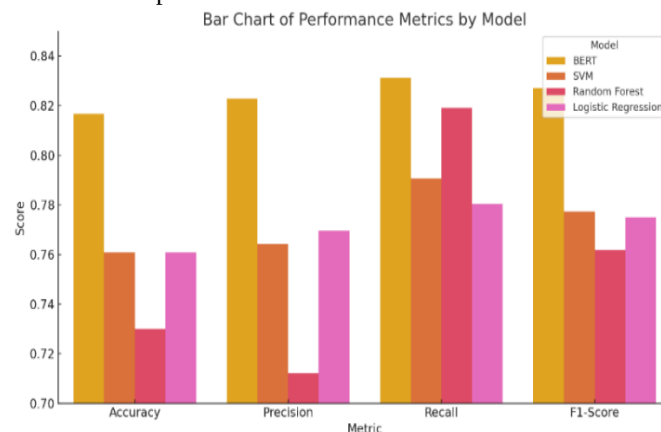


Figure 3. Comparative Bar Chart of Model Performance Metrics

## V. CONCLUSIONS

The findings of this study demonstrate that transformer-based deep learning models, particularly BERT, significantly outperform traditional machine learning approaches in the task of detecting digital harassment in Hindi-language social media content. Leveraging the HASOC Hindi dataset, the BERT model achieved superior performance with an accuracy of 81.67% and an F1-score of 0.827, underscoring its effectiveness in capturing the semantic and contextual nuances of low-resource languages. In comparison, traditional models such as Support Vector Machine and Logistic Regression achieved moderate performance (~76% accuracy), making those viable options for lightweight applications with constrained computational resources. Although Random Forest exhibited relatively high recall for the offensive class (HOF), it suffered from higher false positive rates, indicating class imbalance issues. These results highlight the critical role of advanced preprocessing, tokenization, and contextual modeling in improving classification outcomes, and support the adoption of BERT-based architectures in practical Hindi content moderation systems.

## VI. FUTURE WORK

To build upon the outcomes of this research, several promising directions can be explored. First, the scope of the model can be extended to cover additional Indian languages such as Marathi, Bengali, and Tamil, utilizing the multilingual variants of the HASOC dataset to facilitate cross-lingual transfer learning. Incorporating code-mixed Hindi-English text, which is prevalent in Indian social media discourse, can further enhance the model's robustness and applicability in real-world scenarios.

Furthermore, future work may focus on the deployment of lightweight transformer variants such as DistilBERT or ALBERT, which are more suitable for real-time inference in resource-constrained environments, including mobile and embedded systems. In parallel, critical attention should be given to the ethical dimensions of automated content moderation, particularly in assessing model bias and fairness to ensure equitable treatment across demographic and linguistic groups.

The integration of Explainable AI (XAI) methodologies can enhance system transparency, allowing end-users and moderators to understand and interpret model decisions. Finally, developing a comprehensive deployment pipeline—including APIs, administrative dashboards, and real-time alert mechanisms—will support scalable and actionable digital harassment detection within broader content moderation infrastructures.

## REFERENCES

- [1] Aldhyani T.H.H., Al-Adhaileh M.H. & Alsubari S.N., 2022. *Cyberbullying Identification System Based Deep Learning Algorithms*. Electronics, 11(20), p.3273. Available at: <https://doi.org/10.3390/electronics11203273>.
- [2] Alotaibi M., Alotaibi B. & Razaque A., 2021. *A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media*. Electronics, 10(21), p.2664. Available at: <https://doi.org/10.3390/electronics10212664>.
- [3] Bashar M.A. & Nayak R., 2020. *QutNocturnal@HASOC'19: CNN for hate speech and offensive content identification in Hindi language*. arXiv.org. Available at: <https://doi.org/10.48550/arXiv.2008.12448>.
- [4] Basile V. et al., 2019. *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*. In Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63.
- [5] Bhatia P., Mandal D. & Majumder B., 2021. *One to Rule Them All: Joint Indic Language Hate Speech Detection*. arXiv.org. Available at: <https://doi.org/10.48550/arXiv.2109.13711>.
- [6] Bhardwaj P., Bansal A., Mathur P. & Joshi R., 2020. *Hostility detection dataset in Hindi*. arXiv preprint arXiv:2011.03588. Available at: <https://doi.org/10.48550/arXiv.2011.03588>.
- [7] Deepana M.E., Vigne M., Palanisamy R. & Francis I.S., 2023. *Detection of Cyberbullying Using Machine Learning and Deep Learning*. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), 12(03). Available at: <https://www.ijert.org/detection-of-cyberbullying-using-machine-learning-and-deep-learning>.
- [8] Joshi R., Karnavat R., Jirapure K. & Joshi R., 2021. *Evaluation of Deep Learning Models for Hostility Detection in Hindi Text*. In 2021 6th International Conference for Convergence in Technology (I2CT). Maharashtra, India: IEEE, pp. 1–5. Available at: <https://doi.org/10.1109/I2CT51068.2021.9418073>.
- [9] Narayan R., Patel R. & Joshi R., 2023. *Hate Speech and Offensive Content Detection in Indo-Aryan Languages*. arXiv.org. Available at: <https://doi.org/10.48550/arXiv.2312.05671>.
- [10] Patel J., 2025. *Impact of AI Models on Digital Harassment Prevention*. International Journal of Innovative Research in Technology, 12(1). Available at: <https://ijirt.org/Article?manuscript=181825>.
- [11] Patel J. & Modi N., 2025. *Comprehensive Review of Digital Harassment Prevention and Intervention Strategies*. International Journal for Multidisciplinary Research (IJFMR), 7(1). Available at: <https://doi.org/10.36948/ijfmr.2025.v07i01.35230>.
- [12] Velankar P., Kumar M. & Nigam A., 2021. *Hate and Offensive Speech Detection in Hindi and Marathi using Deep Learning Models*. arXiv.org. Available at: <https://doi.org/10.48550/arXiv.2110.12200>.