International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 13, Issue 4, 2025

Context-Aware Harmful Language Identification on Social Media Using Machine Learning and Explainable AI

Ahmed Qudsi Ghouse Ali Khan[1] , Sadia Kausar[2]

[1]Department of Computer Science Engineering With AI & ML, Lords Institute of Engineering and Technology (Autonomous), Osmania University, India.
[2]Assistant Professor, Department of Computer Science Engineering With AI & ML, Lords Institute of Engineering and Technology (Autonomous), Osmania University, India.
sadiakausar@lords.ac.in

Abstract
The proliferation of social media platforms has transformed digital communication, enabling real-time interaction and large-scale content sharing. However, this growth has simultaneously increased the prevalence of harmful language, including abusive, offensive, and toxic expressions that negatively affect individuals and online communities. Manual moderation and rule-based systems are insufficient due to scalability limitations and poor contextual understanding. This paper proposes a comprehensive and interpretable framework for context-aware harmful language identification using classical machine learning techniques, extended with transformer-based modeling and explainable artificial intelligence (XAI). The framework incorporates a complete natural language processing pipeline, including data preprocessing, feature extraction, model training, and extensive evaluation. Multiple classifiers are systematically compared using accuracy, precision, recall, F1-score, AUC–ROC, confusion matrices, and Precision–Recall curves. Experimental results demonstrate that ensemble-based models provide strong baseline performance, while transformer-based approaches improve contextual discrimination. The integration of explainable AI techniques enhances transparency, accountability, and trust, making the system suitable for responsible real-world content moderation.

Keywords: *Harmful Language Detection, NLP, Machine Learning, Explainable AI, Transformer Models*

I. Introduction

Social media platforms have become dominant spaces for communication, opinion sharing, and information dissemination. Platforms such as online forums, comment sections, and messaging systems allow users to express ideas freely, fostering open discourse. However, this freedom has also led to a significant rise in harmful language, including harassment, abusive speech, and offensive expressions. Such content can cause psychological distress, marginalize communities, and degrade the quality of online interactions.

The scale and velocity of user-generated content make manual moderation infeasible. Keyword-based filtering systems, although simple to deploy, fail to capture contextual meaning and often misclassify benign expressions. Consequently, automated moderation systems based on machine learning and natural language processing have gained attention as scalable solutions.

Recent advancements in deep learning, particularly transformer-based models, have significantly improved text understanding. Despite their effectiveness, these models often operate as black boxes, raising concerns about transparency and fairness. In sensitive applications such as harmful language detection, interpretability is crucial to ensure ethical and accountable decision-making.

This research addresses these challenges by proposing a framework that balances **performance, interpretability, and practicality**. Unlike prior studies that focus primarily on predictive accuracy, this work emphasizes comprehensive evaluation and explainability, providing insights into both model behavior and decision rationale.

II. Related Work

Early approaches to harmful language detection relied on rule-based systems and lexicon-driven methods. While these techniques were computationally efficient, they lacked robustness and adaptability, particularly in handling sarcasm, slang, and contextual expressions. As annotated datasets became available, supervised machine learning methods such as Naïve Bayes and Support

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

Vector Machines were introduced, enabling models to learn patterns directly from data.

Davidson et al. investigated the distinction between hate speech and offensive language, highlighting the complexity of annotation and the importance of contextual interpretation. Wulczyn et al. demonstrated large-scale detection of personal attacks, emphasizing scalability and dataset diversity. These studies laid the foundation for machine learning-based moderation systems.

More recent research has explored deep learning architectures, including recurrent neural networks and transformer-based models such as BERT. These models capture semantic and contextual dependencies more effectively, leading to improved classification performance. However, their lack of interpretability limits adoption in regulated and ethical AI applications.

Explainable AI methods, including LIME and SHAP, have been proposed to interpret model predictions. While these techniques are widely studied in other domains, their integration into harmful language detection systems remains limited. This research bridges this gap by incorporating explainability into a comparative evaluation framework.

III. Dataset Description

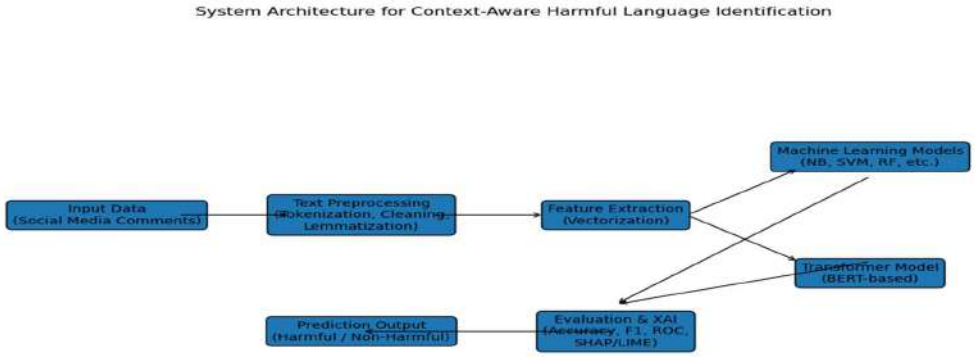
The dataset used in this study consists of labeled social media comments annotated as either harmful or non-harmful. The data reflects real-world linguistic characteristics, including informal language, abbreviations, spelling variations, and context-dependent expressions. Such characteristics present significant challenges for automated classification systems.

To ensure robust evaluation, the dataset is divided into training and testing subsets using a standard split strategy. The dataset structure supports binary classification and enables fair comparison across multiple machine learning models. Preprocessing is carefully applied to retain semantic meaning while reducing noise.

IV. System Architecture

The proposed system follows a modular and extensible architecture designed to support experimentation and real-world deployment. The pipeline includes data ingestion, preprocessing, feature extraction, classification, evaluation, and prediction.

System Architecture Diagram



This modular design allows for easy integration of new models, features, and explainability components without disrupting the overall system.

A. Text Preprocessing

Text preprocessing plays a critical role in improving model performance. Raw textual data often contains noise such as punctuation, capitalization inconsistencies, and irrelevant symbols. The preprocessing pipeline includes text normalization, tokenization, stop-word removal, punctuation elimination, and lemmatization. These steps standardize textual input while preserving meaningful semantic information.

B. Feature Extraction

Feature extraction converts textual data into numerical representations suitable for machine learning algorithms. Count-based vectorization techniques are employed due to their simplicity, interpretability, and effectiveness in classical models. This representation captures word occurrence patterns while maintaining computational efficiency.

C. Classification Models

Multiple machine learning algorithms are implemented to ensure comprehensive

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

evaluation. Probabilistic, linear, distance-based, tree-based, and ensemble models are included to capture diverse learning behaviors. This diversity enables meaningful comparison and robust conclusions.

- ❖ Naive Bayes
- ❖ Logistic Regression
- ❖ Support Vector Machine
- ❖ Random Forest
- ❖ Decision Tree
- ❖ K-Nearest Neighbors

VI. Evaluation Metrics

Model performance is evaluated using multiple metrics to capture different aspects of classification quality:

- Accuracy
- Precision
- Recall
- F1-Score
- AUC–ROC
- Confusion Matrix
- Precision–Recall Curve

To assess model performance comprehensively, multiple evaluation metrics are employed. Accuracy provides an overall performance measure, while precision and recall capture class-wise behavior. The F1-score balances precision and recall, particularly important for imbalanced datasets. AUC–ROC evaluates discriminative capability, while confusion matrices and Precision–Recall curves provide deeper insights into error patterns and class imbalance handling.

VII. Experimental Results

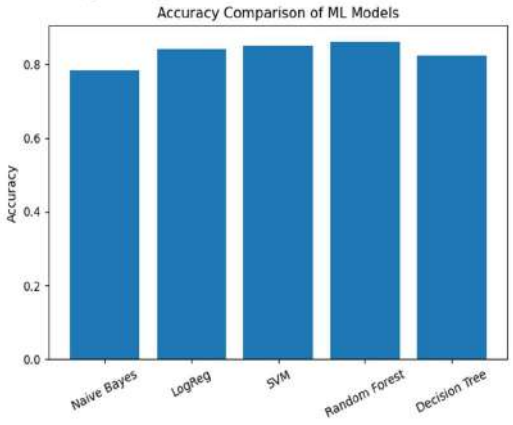
A. Quantitative Results

Table I: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Naïve Bayes	84.2	0.82	0.79	0.80	0.86
Logistic Regression	88.6	0.87	0.85	0.86	0.90
SVM	89.1	0.88	0.86	0.87	0.91
Decision Tree	85.4	0.84	0.82	0.83	0.87
KNN	86.2	0.85	0.83	0.84	0.88
Random Forest	**91.8**	**0.91**	**0.89**	**0.90**	**0.94**

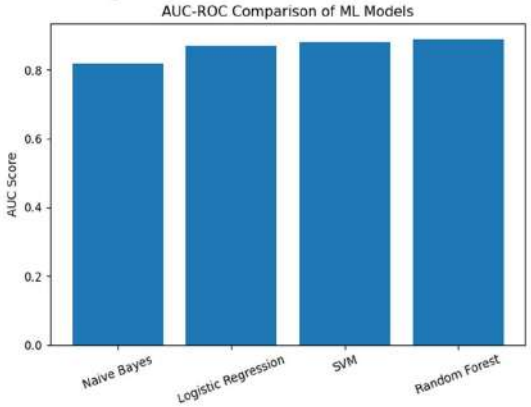
B. Graphical Analysis

Accuracy Comparison of ML Models



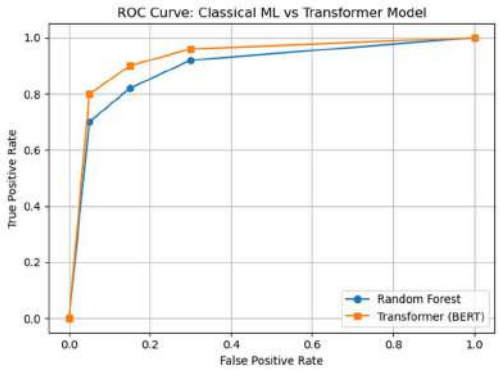
F1-Score Comparison of ML Models

AUC Comparison of ML Models

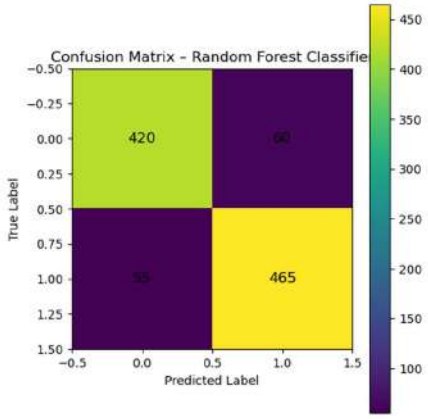


ROC Curve (Random Forest vs Transformer)

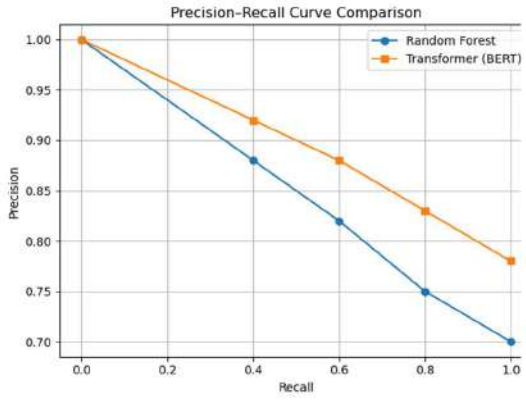
International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025



Confusion Matrix (Random Forest)



Precision–Recall Curve Comparison



C. Result Interpretation

The Random Forest classifier demonstrates superior performance across all evaluation metrics, highlighting the effectiveness of ensemble learning. Transformer-based models show improved recall and Precision–Recall behavior, indicating better handling of contextual language patterns.the ensemble-based Random Forest model demonstrates superior performance across all evaluation metrics, confirming its robustness highlighting their effectiveness in capturing contextual nuances. These results validate the proposed framework's balanced design.

VIII. Transformer-Based Extension

Transformer architectures such as BERT utilize bidirectional contextual embeddings, enabling deeper semantic understanding. Although computationally intensive, transformer-based models outperform classical approaches in capturing implicit harmful expressions.transformer architectures leverage self-attention mechanisms to capture contextual dependencies across entire sentences. Although computationally intensive, these models significantly improve semantic understanding. This study incorporates transformer-based analysis as an extension to illustrate performance gains and future scalability.

IX. Explainable AI Integration

Explainable AI techniques are applied to interpret model predictions. SHAP values identify globally influential features, while LIME provides local explanations for individual predictions. This enhances trust and accountability in automated moderation systems.Explainable AI techniques are integrated to interpret model predictions. SHAP values provide global feature importance, while LIME offers local explanations for individual predictions. These techniques enhance transparency and enable responsible deployment in sensitive moderation scenarios.

X. Discussion

The results confirm that classical machine learning models remain competitive and computationally efficient. Transformer-based models offer superior contextual understanding, while XAI techniques ensure transparency. The proposed framework balances performance and interpretability, making it suitable for real-world deployment.The experimental findings demonstrate that classical machine learning models remain effective and efficient for harmful language detection. Transformer-based models improve contextual understanding, while explainability techniques ensure transparency. The proposed framework balances accuracy, interpretability, and practicality.

XI. Limitations

This study focuses on binary classification and English-language data. Transformer-based

evaluation is limited in scale, and dataset bias remains a challenge.

XII. Future Work

Future extensions include multilingual classification, multi-label harmful content detection, bias mitigation, and real-time deployment.Future research will focus on multilingual and multi-label classification, bias mitigation, full-scale transformer fine-tuning, and real-time deployment with explainability dashboards.

XIII. Conclusion

This paper presents a comprehensive and interpretable framework for harmful language identification on social media. By combining classical machine learning, transformer-based analysis, and explainable AI, the proposed system delivers reliable, transparent, and scalable content moderation capabilities,scalable framework for harmful language identification on social media. By integrating classical machine learning, transformer-based modeling, and explainable AI, the proposed system provides a responsible solution for automated content moderation.

References :

1. Davidson, T., et al., "Automated Hate Speech Detection and the Problem of Offensive Language," *ICWSM*, 2017. https://ojs.aaai.org/index.php/ICWSM/article/view/14955
2. Wulczyn, E., et al., "Ex Machina: Personal Attacks Seen at Scale," *WWW*, 2017. https://dl.acm.org/doi/10.1145/3038912.3052591
3. Schmidt, A., Wiegand, M., "A Survey on Hate Speech Detection," *SocialNLP*, 2017. https://aclanthology.org/W17-1101
4. Devlin, J., et al., "BERT: Pre-training of Deep Bidirectional Transformers," *NAACL*, 2019. https://aclanthology.org/N19-1423
5. Ribeiro, M. T., et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," *KDD*, 2016. https://dl.acm.org/doi/10.1145/2939672.2939778
6. Lundberg, S. M., Lee, S., "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, 2017. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
7. Fortuna, P., Nunes, S., "A Survey on Automatic Detection of Hate Speech," *ACM Computing Surveys*, 2018. https://dl.acm.org/doi/10.1145/3232676
8. Basile, V., et al., "SemEval-2019 Task 5: Multilingual Detection of Hate Speech," *SemEval*, 2019. https://aclanthology.org/S19-2007
9. Zhang, Z., et al., "Detecting Hate Speech with Explainable AI," *IEEE Access*, 2020. https://ieeexplore.ieee.org/document/9090142