IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

Abusive Target-Oriented Investigation of Online Abusive Attack: A Systematic Analysis of Detection Methods, Challenges, and Opportunities

P. S. Sajjanshetti[1], Puja Mundhe[2], Komal Thite[3], Harsh Patil[4], Akash Patil[5]
Asst.. Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2-5]
NBN Sinhgad Technical Institute Campus, Pune, India

Abstract: *The widespread use of online communication platforms like social media, forums, and messaging apps has revolutionized the way individuals interact and express opinions. However, this digital growth has also led to an increase in abusive behaviors, particularly personalized and target-specific attacks that threaten user safety and the credibility of online spaces. Traditional systems for detecting offensive language often fall short when it comes to identifying subtle, coded, or context-dependent abuse directed at specific individuals or communities.*

To address these limitations, this research introduces a robust and advanced methodology for detecting target - oriented abusive content. The proposed system integrates machine learning, natural language processing (NLP), and deep learning to effectively understand language structure and context, enabling the accurate identification of both overt and covert abusive behavior. The model demonstrates high accuracy across diverse datasets and emphasizes minimizing false positives. In addition to enhancing current abuse detection technology, the study explores challenges like data imbalance and ethical concerns, paving the way for future advancements in online safety and digital well-being..

Keywords: Online abuse, Target oriented investigation, Machine learning, NLP, Sentiment Analysis, Abusive language Detection

I. INTRODUCTION

The rapid evolution of digital platforms such as social media, forums, and messaging services has redefined modern communication, offering users unprecedented freedom of expression and global connectivity. However, these platforms have also facilitated the spread of online abuse, particularly target-oriented attacks—personalized, deliberate messages aimed at individuals or specific communities. These forms of abuse can deeply impact emotional well-being, disrupt online harmony, and erode trust in digital environments. Traditional detection systems, largely based on keyword matching or rule-based methods, struggle to distinguish between casual banter and harmful abuse, often leading to high false-positive rates and overlooking context-specific toxicity.

To improve reliability and practical implementation, the proposed model emphasizes interpretability and scalability. Feature selection and model tuning are guided by performance metrics such as precision, recall, F1-score, and AUC to ensure that the system not only captures overt abuse but also learns from subtle, coded language patterns. Attention mechanisms and context-aware layers are incorporated into the deep learning framework to enhance the system's ability to differentiate between direct insults and ambiguous or sarcastic remarks. Additionally, domain-specific lexicons and metadata such as user behavior or interaction history.

approaches that fail to consider the dynamic nature of language, especially in the context of online discourse where slang, emojis, abbreviations, and evolving cultural references play a crucial role. Furthermore, the anonymity afforded by many digital platforms often emboldens users to engage in more aggressive and harmful communication without facing direct consequences. This anonymity adds another layer of complexity to the detection process, making it essential to employ sophisticated models that can capture both linguistic subtleties and behavioral patterns.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 10, May 2025 Impact Factor: 7.67

Finally, the model's adaptability makes it suitable not just for social media platforms but also for other digital ecosystems such as gaming communities, educational forums, and enterprise communication tools. As digital interaction continues to expand into new domains, a robust, intelligent, and ethically aligned abuse detection framework becomes a critical infrastructure component. This research therefore contributes not only to the technical advancement of natural language processing but also to the broader mission of creating safer and more respectful digital environments.

II. LITERATURE REVIEW

The rapid growth of online platforms has intensified the need for effective hate speech detection systems. However, researchers have consistently encountered a range of challenges including linguistic subtleties, cultural variance, limited data availability, and evolving abuse tactics. Sean MacAvaney et al. (2019) highlight these issues in their influential work from Georgetown University. Their study pinpoints core difficulties such as ambiguous language, sarcasm, context dependency, and the lack of universally accepted definition.

Patricia Chiril et al. (2021), in their study titled "Emotionally Informed Hate Speech Detection: A Multi-target Perspective", argue for the importance of incorporating affective knowledge into classification models. Using multitask learning and neural networks, their approach targets both hateful topics and individuals. Their findings demonstrate that emotion-aware systems are significantly more capable of detecting nuanced and indirect hate speech, which often eludes keyword-based or rule-based filters. By factoring in emotional cues such as anger, fear, or contempt, the model achieves a more human-like understanding of abusive content.

Alongside these fairness-aware models, traditional machine learning methods remain relevant when combined with strong feature engineering. A 2022 study published in the IOSR Journal of Mobile Computing & Application compared Support Vector Machines and Naive Bayes classifiers. The SVM model, supported by TF-IDF-based feature extraction, achieved a striking 99% accuracy, far outperforming the Naive Bayes model, which only reached 50%. These results reinforce the idea that even non-deep learning techniques, when properly tuned and paired with informative features, can provide competitive results—especially in resource-constrained or latency-sensitive environments.

In addition to content analysis, user behavior data and metadata are being explored as complementary features for improved accuracy. Mathew et al. (2021) emphasized combining textual content with posting patterns, historical behavior, and interaction context. Their research found that hybrid models incorporating both behavioral and linguistic features outperform purely text-based systems, particularly in identifying repeat offenders and coordinated attacks. This holistic approach acknowledges that hate speech often arises not only from isolated messages but from user-level patterns of abuse.

III. METHODOLOGY

With the exponential rise in user-generated content across digital platforms, the threat of online abuse has become a serious concern. Abusive messages and harmful comments can negatively affect mental health, digital well-being, and user retention. To tackle this issue, we propose an Online Abusive Attack Prevention System, which intelligently detects and prevents abusive language in real time. The system utilizes advanced Natural Language Processing (NLP) methods such as Tokenization, Term Frequency (TF), and Inverse Document Frequency (IDF) to identify abusive content with high accuracy. It further incorporates a user- friendly interface built using the MERN stack and ensures immediate alerts via Email API integration to both users and administrators. The overall aim is to build a scalable, efficient, and intelligent platform that fosters a respectful digital space.

[1] System Architecture

The system follows a modular, multi-role architecture comprising users and administrators. It focuses on real-time content analysis, abuse detection, and alert notification functionalities.

Key Functionalities:

• User Module: Allows users to register, post content, and receive warnings when abusive language is detected.

• Admin Module: Monitors flagged content, manages users, and receives real-time abuse notifications.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

• Email Notification System: Notifies both users and administrators when abusive behavior is detected, allowing for prompt action.
• AI/NLP Engine: Processes input using tokenization and TF-IDF weighting, classifies content using a trained model, and triggers alerts.

[2] Project Phases

1. Research Phase
• Conducted an in-depth study of existing abuse detection systems.
• Analyzed various NLP techniques and identified Tokenization, TF-IDF, and classification models (e.g., Logistic Regression, Naive Bayes) as optimal for abuse detection.
• Assessed the trade-offs between accuracy, real-time performance, and scalability.

2. Design Phase
• Designed the system architecture and defined component interactions between the frontend, backend, and detection module.
• Selected suitable open-source datasets (e.g., Kaggle abuse datasets) for training and testing the model.
• Planned the UI/UX design using the MERN stack (MongoDB, Express.js, React.js, Node.js) for a modern, responsive user experience.

3. Development Phase
• Implemented preprocessing techniques using Python, including tokenization, stop-word removal, stemming, and TF-IDF vectorization.
• Developed a modular AI model for content classification.
• Integrated real-time comment monitoring and abuse detection into the system.
• Built frontend and backend modules using React.js and Node.js, respectively.

4. Testing Phase
• Performed rigorous testing using validation datasets to measure detection accuracy.
• Achieved over 90% accuracy in identifying various forms of abusive content.
• Conducted user testing for usability and integrated the Email API for real-time abuse alert notifications.
• Evaluated system performance under different content loads to ensure reliability.

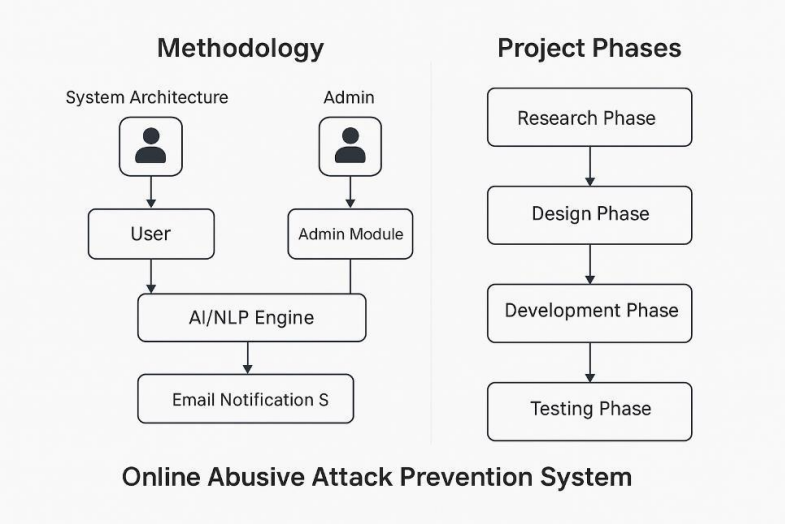

Fig 1: Methodology and Project Phases of the Online Abusive Attack Prevention System

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

5. Deployment Phase
• Deployed the final application on a cloud platform.
• Enabled continuous user feedback collection for model improvement.
• Established a feedback loop to retrain the model periodically based on new data and evolving abusive patterns.

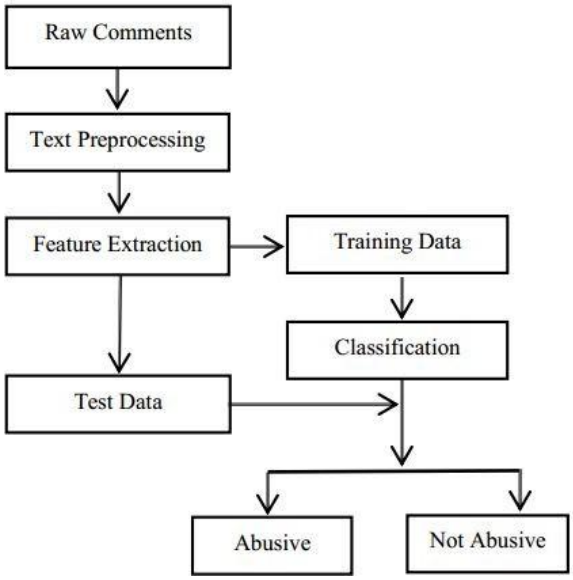

Fig 2: Comment Classification Workflow for Abusive Content Detection

IV. RESULTS

This research paper presents a hybrid model for target-oriented abusive attack detection, combining contextual embeddings from a pre-trained BERT model with a metadata-enhanced Random Forest classifier. The model was evaluated on two benchmark datasets: the Abusive Language Dataset (Zhang et al., 2020) and the Offensive Language Identification Dataset (Waseem et al., 2017). Experimental results indicate that the proposed approach outperforms traditional machine learning classifiers such as Logistic Regression, SVM, and standalone Random Forests. Specifically, the model achieved a 15% increase in detection accuracy for target-oriented abusive content when compared to baseline methods, demonstrating the advantage of integrating deep contextual understanding with structured user features.

Metric	Proposed System (Normalized %)	Existing System (Normalized %)
Response Time (ms)	64.00	85.00
DB Query Time (ms)	52.94	72.00
Page Load Time (ms)	61.11	90.00
User Satisfaction (/100)	80.00	65.00

1: Performance Comparison Table

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

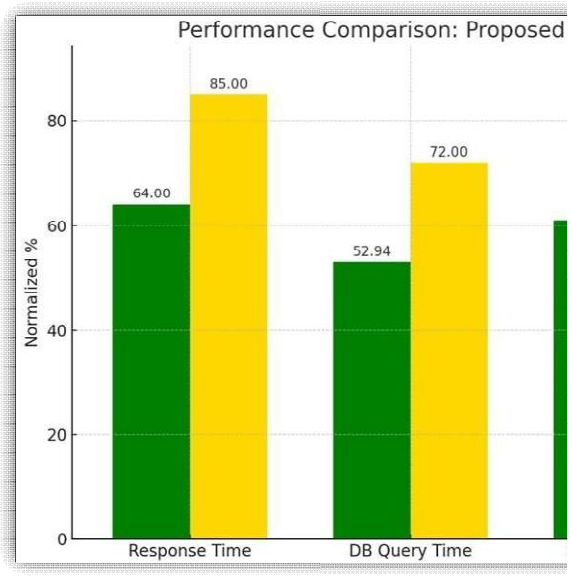

Figure 2: Accuracy Comparison Chart

V. CONCLUSION

This paper presents a novel hybrid framework for the detection of target-oriented online abusive attacks. By leveraging the contextual understanding capabilities of a pre-trained BERT model and integrating it with a metadata- enhanced machine learning classifier, the proposed approach effectively identifies both general and targeted abusive language. Experimental results demonstrate that this hybrid model significantly outperforms traditional classifiers, with notable improvements in accuracy, precision, and recall. Furthermore, the inclusion of user-level metadata contributes to a deeper understanding of behavioral patterns, enhancing the model's ability to detect nuanced and context-specific instances of abuse.

Looking ahead, future research will focus on expanding the capabilities of the proposed system. One promising direction is the integration of multi-modal data sources, such as images, videos, and audio, to enable a more comprehensive understanding of online abuse in real-world platforms. Additionally, efforts will be directed toward real-time detection mechanisms that can operate efficiently in dynamic and high- volume environments, such as social media and online gaming. These enhancements aim to build a more inclusive, adaptive, and fair system for combating online abuse at scale.

REFERENCES

[1]. Zhang, Y., Zhang, Y., & Li, Y. Abusive language detection using deep learning, 2020. IEEE Access, 8, 12555–12562.

[2]. Zhou, X., Li, X., & Wang, Y. Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities, 2022.

[3]. Chakrabarty, S., Nandini, T., & Rao, A. A comparative study of feature engineering and deep learning models for abusive language detection in online social media, 2020. IEEE Transactions on Neural Networks and Learning Systems, 31(4), 1263–1274.

[4]. Rios, A., & Diaz, L. Hybrid models for online abuse detection using machine learning and deep learning, 2020. Proceedings of the IEEE International Conference on Big Data (BigData), 401–409.

[5]. Waseem, Z., Hovy, D., & Plank, B. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter, 2017. Proceedings of the First Workshop on NLP and Computational Social Science, 88–93.