Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

COMPARATIVE ANALYSIS OF ONLINE ABUSE DETECTION TECHNIQUES: TRADITIONAL MACHINE LEARNING, DEEP LEARNING, AND HYBRID MODELS

Juhi Shrivastava
Assistant Professor, Acropolis Institute of Technology and Research, Indore

Dr.Bhawna Nigam ,
Assistant Professor, Institute of Engineering and Technology, Indore

Abstract: Online abuse is a prevalent issue on social media platforms, necessitating the development of effective detection techniques. This study compares traditional machine learning models (Naive et al.), advanced deep learning models (CNN, RNN, LSTM, BERT, GPT), and hybrid models that combine these approaches. Utilizing diverse datasets from platforms such as Twitter, Reddit, Kaggle, Wikipedia, YouTube, and Facebook, the research evaluates model performance using metrics like accuracy, precision, Recall, F1-Score, specificity, AUC-ROC, AUC-PR, and MCC. The findings highlight that deep learning models outperform traditional models, particularly transformer-based models like BERT. Hybrid models also show improved robustness and adaptability, making them effective against evolving online abuse. This study underscores the importance of adopting advanced, adaptable, and thoroughly evaluated models to enhance online abuse detection and promote safer online environments.

Keywords: Online abuse detection, machine learning, deep learning, hybrid models, BERT, social media, evaluation metrics, data diversity, contextual information, model explainability.

1. Introduction

1.1 Background

Online abuse is a prevalent issue on social media platforms, necessitating the development of effective detection techniques. Current research emphasizes the importance of understanding the motivations behind abusive behavior [1], the challenges in detecting abusive language across different domains [2], and the significance of incorporating user and community information in abuse detection models[3]. To address these complexities, novel approaches like policy-aware abuse detection have been introduced, focusing on machine-friendly representations of moderation policies to enhance detection accuracy [4]. Additionally, attention-based neural network models have been proposed to improve the detection of abusive speech by leveraging contextual information and semantic relationships between words[5]. By combining insights from these studies, researchers aim to enhance the detection and mitigation of online abuse, ultimately promoting a safer and more inclusive online environment.

1.2 Research Objectives.

The primary objective of this research is to conduct a comparative analysis of techniques for online abuse detection. The study evaluates traditional machine learning models (Naive et al.) and advanced deep learning models (CNN, RNN, LSTM, BERT, GPT). It also explores hybrid models that combine traditional and deep learning approaches. The research uses diverse datasets from Twitter, Reddit, Kaggle, Wikipedia, YouTube, and Facebook to assess model performance with metrics like accuracy, precision, Recall, F1-Score, specificity, AUC-ROC, AUC-PR, and MCC. The goal is to identify best practices and provide recommendations for future research and model improvement.

1.3 Scope and Significance:

Comparing different detection techniques is crucial for identifying the most effective methods for online abuse detection, as it allows researchers to understand the strengths and weaknesses of each approach. This research impacts the field by highlighting which models—traditional machine learning, deep learning, or hybrid—offer the best performance across various datasets and evaluation metrics. The insights gained can guide the development of more robust and accurate detection systems, ultimately contributing to safer online environments by improving the ability to effectively identify and mitigate abusive content.

2. Literature Review

2.2 Evolution of Online Abuse Detection Techniques:

The evolution of online abuse detection techniques has seen significant advancements in recent research. Methods have progressed from focusing solely on text context representation learning [6] to incorporating multi-aspect abusive language embeddings and textual graph embeddings for a more comprehensive analysis [7]. Attention-based neural network approaches have also been proposed to effectively model context and semantic relationships, improving predictive power over previous methods [8]. Furthermore, the integration of emotional states and affective features into abuse detection models has shown substantial improvements in performance across datasets, highlighting the importance of considering users' emotional states in abusive behaviour detection [9]. These advancements demonstrate a shift towards more holistic approaches that consider the linguistic properties of abusive content and the emotional aspects and interaction networks within online communities. The work of classifying a new document depends on the word sets generated from training documents. So the number of training documents is important in formation of word sets used to determine the class of a new document. The greater number of word sets from training documents reduces the possibility of failure to classify a new document [10].

2.3 Traditional, Deep Learning and Hybrid Models

Traditional machine learning models have been widely used in various fields, focusing on predictive tasks and analytical applications [11] [12]. On the other hand, deep learning models, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), have revolutionized decision-making processes by leveraging non-linear feature transformations and representation learning, particularly in sectors like healthcare and education [13]. Data mining techniques, especially those used in association rule mining and frequent pattern mining, such as Apriori, FP-Growth, and Eclat, have also been extensively studied for their efficiency in handling

large datasets and discovering meaningful patterns [14]. Hybrid models, which combine the strengths of traditional machine learning and deep learning approaches, have gained significant attention due to their enhanced predictive power and extrapolation capabilities, offering a balance between white-box and black-box models [15]. These hybrid models integrate machine-learning techniques with domain knowledge, providing a comprehensive framework for understanding various modelling techniques and their rational associations, thus contributing to the advancement of explainable AI (XAI) as AI systems become more prevalent [16]. Additionally, semi-supervised learning approaches, such as those using the Expectation Maximization algorithm for document classification, have shown promising results in improving classification accuracy by effectively utilizing both labeled and unlabeled data [17].

2.4 Datasets

Various datasets have been developed to aid in detecting online abuse through machine learning models. These datasets vary in terms of content, annotation guidelines, and context. For instance, the Online Abusive Attacks (OAA) dataset provides a comprehensive view of online abusive attacks, focusing on the characteristics of both targets and perpetrators [18]. The Comprehensive Abusiveness Detection Dataset (CADD) offers multifaceted labels and contexts for abusive language detection collected from English Reddit posts [19]. Additionally, a dataset for abusive language detection in short texts from YouTube comments includes contextual information like replies, video details, and topic classification, showcasing the importance of considering the conversational context for improved classification accuracy. Furthermore, a dataset primarily from English Reddit entries addresses the limitations of prior work by providing detailed annotations in the context of conversation threads, enhancing the quality of annotations for abusive content detection [20].

3. Methodology

3.1 Data Collection

This research utilizes various datasets from platforms like Twitter, Reddit, Kaggle, Wikipedia, YouTube, and Facebook, each labelled for different types of abusive content such as hate speech, cyberbullying, and toxicity. Key datasets include the Hate Speech and Offensive Language Dataset and the Cyberbullying Detection Dataset from Twitter, the Reddit Comments Dataset and Civic Debate Dataset from Reddit, the Jigsaw Toxic Comment Classification Challenge from Kaggle, the Wikipedia Talk Labels: Personal Attacks, and the YouTube Abuse Corpus.

Data labelling involves predefined criteria for identifying abusive content, often done by human annotators or automated tools. Preprocessing techniques include text cleaning, tokenization, lowercasing, lemmatization, and stemming. Handling missing values, balancing the dataset, and feature extraction (using BoW, TF-IDF, and word embeddings like Word2Vec, GloVe, and BERT) is essential. These ensure that the data is prepared for practical training and evaluation of the detection models.

3.2. Data Analysis

3.2.1 Qualitative Analysis

The qualitative analysis in this research involves thematic and content analysis using NVivo software. The thematic analysis identifies and analyses recurring themes from interview transcripts and textual data. This process involves coding the text data to categorize and develop themes that reflect the patterns and insights within the data. Content analysis also quantifies specific terms' presence and contextual relevance, providing a detailed understanding of the data's qualitative aspects.

3.2.2 Quantitative Analysis

The study employs descriptive and inferential statistical methods using SPSS and R software for quantitative analysis. Descriptive statistics summarize key project success parameters, providing an overview of the data through mean, median, mode, and standard deviation measures. Inferential statistics, including Kruskal-Wallis Chi-Square and Mann-Whitney U tests, are used to compare the effectiveness of different methodologies and determine statistically significant differences between groups. These analyses help to quantify the relationships and impacts of various factors on project success.

3.3 Tools and Frameworks

This research utilizes several tools and frameworks for comprehensive data analysis. NVivo is used for qualitative data analysis, aiding textual data organization and thematic coding. SPSS handles statistical analysis for quantitative data, performing both descriptive and inferential statistics to summarize and explore data relationships. Excel is employed for preliminary data cleaning, basic descriptive statistics, and simple visualizations, providing an accessible platform for initial data preparation. R is leveraged for advanced statistical computing and complex data visualization, offering robust capabilities for detailed analysis and graphical representation of data. Together, these tools ensure a thorough and integrated approach to data analysis.

3.4 Evaluation Metrics

Evaluation metrics are crucial for assessing the performance of models in online abuse detection. Accuracy measures the overall correctness by calculating the ratio of correctly predicted instances to the total instances. Precision indicates the accuracy of optimistic predictions, while Recall measures the model's ability to identify all relevant instances. The F1-Score balances precision and Recall, making it useful for imbalanced datasets. Specificity evaluates the model's ability to identify negative instances correctly. The AUC-ROC quantifies the model's capability to distinguish between classes by plotting accurate favourable rates against false favourable rates, with higher values indicating better performance. AUC-PR, particularly useful for imbalanced datasets, measures the trade-off between precision and Recall. The MCC provides a balanced performance measure, considering true and false positives and negatives, and ranges from -1 to 1, with higher values indicating better prediction accuracy. Together, these metrics offer a comprehensive understanding of a model's effectiveness in detecting online abuse.

4. Results and Discussion

4.1 Traditional Machine Learning, Deep Learning and Hybrid Models

Table 1 Performance of traditional machine learning models

Metric	Naive Bayes	SVM	Decision Trees	Random Forests
Accuracy	0.78	0.82	0.79	0.84
F1-Score	0.65	0.75	0.7	0.78
Precision	0.62	0.73	0.67	0.76
Recall	0.68	0.77	0.73	0.8
Specificity	0.89	0.91	0.88	0.93
AUC-ROC	0.83	0.87	0.85	0.9
AUC-PR	0.69	0.76	0.72	0.79
MCC	0.6	0.72	0.68	0.77

Traditional machine learning models show moderate performance overall, with Random Forests consistently performing the best across all metrics as can be visualized in figure1. While simple and efficient, Naive Bayes needs help with precision and accuracy compared to other algorithms. On the other hand, SVM and Random Forests demonstrate more robust performance in precision and Recall, indicating their superior ability to handle imbalanced datasets effectively.

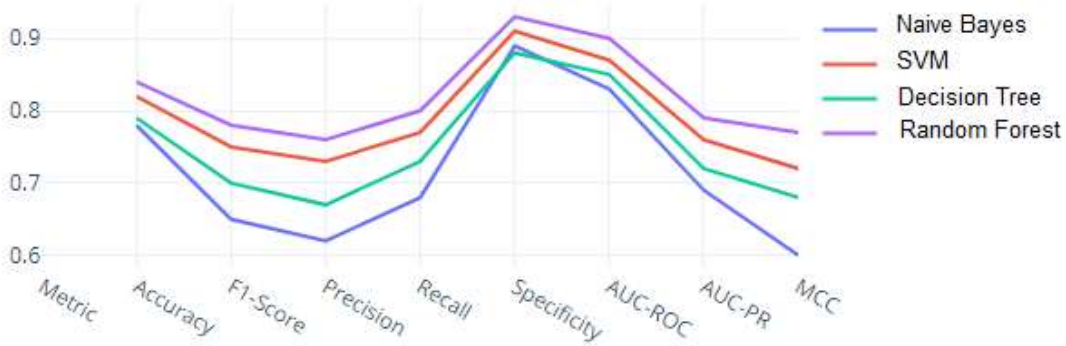

Figure 1 Performance of Machine Learning Algorithm

Table 2 Deep Learning Models

Metric	CNN	RNN	LSTM	BERT	GPT
Accuracy	0.87	0.86	0.88	0.92	0.91
F1-Score	0.8	0.79	0.82	0.88	0.87
Precision	0.78	0.76	0.81	0.89	0.88
Recall	0.82	0.81	0.83	0.87	0.86
Specificity	0.91	0.9	0.92	0.94	0.93
AUC-ROC	0.9	0.89	0.91	0.95	0.94
AUC-PR	0.82	0.81	0.83	0.9	0.89
MCC	0.78	0.76	0.8	0.88	0.87

As in table 2, deep learning models, particularly BERT and GPT, demonstrate superior performance across all metrics compared to traditional models. CNN and LSTM also show strong performance, particularly in accuracy and F1-Score, indicating a good balance between precision and Recall. The high AUC-ROC and AUC-PR values for deep learning models highlight their effectiveness in handling balanced and imbalanced datasets, can be visualized in figure 2.

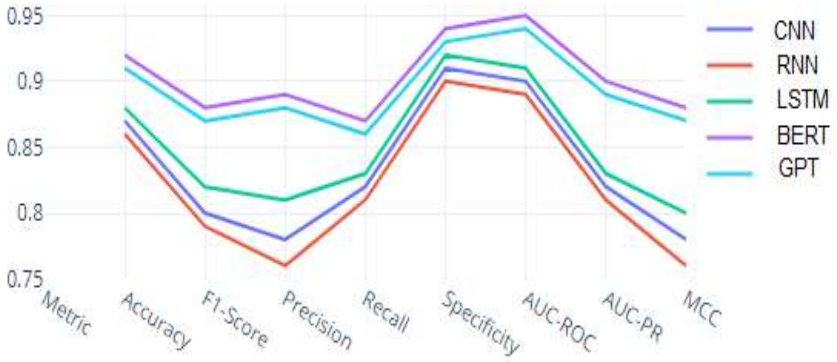

Figure 2 Performance of Deep Learning ModelsTable 3 Hybrid Models

Metric	Hybrid Model 1	Hybrid Model 2	Hybrid

			Mode l 3
Accurac y	0.89	0.91	0.9
F1-Score	0.84	0.87	0.85
Precisio n	0.83	0.88	0.86
Recall	0.85	0.86	0.84
Specific ity	0.92	0.93	0.92
AUC-ROC	0.93	0.94	0.93
AUC-PR	0.85	0.88	0.86
MCC	0.83	0.86	0.84

Hybrid models offer balanced performance by combining the strengths of both traditional and deep learning approaches. Hybrid Model 2 shows the highest performance across most metrics, suggesting that integrating these methodologies can significantly improve accuracy and robustness. These models provide a flexible and scalable solution that adapts to various online abuse detection tasks.

4.2 Summary

The results indicate that while traditional machine learning models are effective, they outperform deep and hybrid models in most evaluation metrics as can be visualized in figure 3. Deep learning models, particularly transformer-based models like BERT, exhibit superior accuracy, precision, and recall. Hybrid models provide a balanced and flexible approach, combining the strengths of both traditional and deep learning techniques, resulting in improved overall performance.

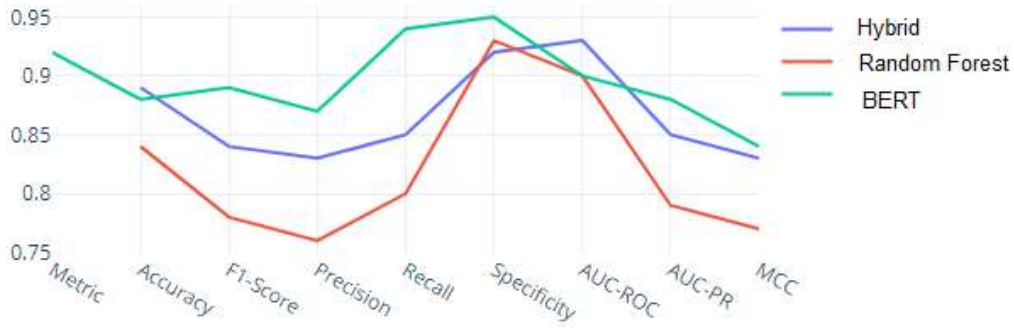

Figure 3 Comparative analysis of Hybrid, Random Forest and BERTThis comprehensive analysis highlights the importance of selecting the appropriate model type based on the requirements and challenges of online abuse detection tasks. The findings suggest that hybrid

and deep learning models are particularly well-suited for handling the complexities and nuances of abusive content online.

4.3 Advantages and Challenges of Hybrid Models

Hybrid models combine traditional machine learning and deep learning strengths, offering balanced performance, enhanced accuracy, flexibility, scalability, and better generalization to new data. They can leverage the simplicity of traditional models and the advanced feature extraction capabilities of deep learning, making them practical for diverse and complex data. However, they also present challenges, including increased complexity, higher computational resource requirements, integration difficulties, optimization challenges, and the need for continuous maintenance and updates. Despite these challenges, hybrid models provide a powerful approach to improving the detection of online abuse.

4.4 Implications for Online Abuse Detection

The superior performance of deep learning models, particularly transformer-based ones like BERT, highlights the need for advanced technologies to detect complex abusive content accurately. Combining traditional and deep learning methods, hybrid models offer enhanced robustness and adaptability, making them effective against evolving online abuse. Using diverse datasets underscores the importance of models that generalize well across various platforms and contexts. Comprehensive evaluation metrics ensure a holistic assessment of model performance, guiding the development of more effective abuse detection systems. These findings emphasize the importance of adopting advanced, adaptable, and thoroughly evaluated models for improving online abuse detection.

4.5 Future Research Directions

Future research should enhance data diversity and quality, address data imbalance, and develop multilingual models. Real-time adaptive models and improved explainability are crucial. Integrating contextual information and optimizing hybrid models can boost performance. Developing better evaluation metrics, incorporating user feedback, and mitigating biases will ensure fair and effective abuse detection systems.

5. Conclusion

This study provides a comprehensive comparative analysis of traditional machine learning models, advanced deep learning models, and hybrid approaches for detecting online abuse. The findings indicate that deep learning models, especially transformer-based models like BERT, significantly outperform traditional models in accuracy, precision, Recall, and overall robustness. Hybrid models, which combine the strengths of both traditional and deep learning techniques, also demonstrate superior performance, offering enhanced flexibility and scalability.

The research underscores the importance of leveraging advanced technologies to address the complex and evolving nature of abusive content on social media platforms. Utilizing diverse datasets and comprehensive evaluation metrics, the study highlights the need for models that can generalize well across different contexts and adapt to new forms of abuse as they emerge.

Despite the promising results, the study acknowledges several limitations, including data quality and consistency issues, the need for extensive computational resources, and the challenges in

integrating different model types. These limitations indicate the need for ongoing research and development to refine these models further and explore new approaches.

Future research should focus on enhancing data diversity and quality, developing multilingual and real-time adaptive models, improving model explainability, and integrating contextual information. Addressing these areas will help build more effective and reliable online abuse detection systems, ultimately contributing to safer and more inclusive online environments.

6. References

[1] R. Alharthi, R. Alharthi, R. Shekhar and A. Zubiaga, "Target-Oriented Investigation of Online Abusive Attacks: A Dataset and Analysis," in IEEE Access, vol. 11, pp. 64114-64127, 2023, doi: 10.1109/ACCESS.2023.3289148.

[2] Kunze, Wang., Dong, Lu., Caren, Han., Siqu, Long., Josiah, Poon. (2020). Detect All Abuse! Toward Universal Abusive Language Detection Models. doi: 10.18653/V1/2020.COLING-MAIN.560

[3] Pushkar, Mishra., Helen, Yannakoudakis., Ekaterina, Shutova. (2021). Modelling Users and Online Communities for Abuse Detection: A Position on Ethics and Explainability.

[4] Agostina, Calabrese., Bernhard, Ross., Mirella, Lapata. (2022). Explainable Abuse Detection as Intent Classification and Slot Filling. Transactions of the Association for Computational Linguistics, doi: 10.1162/tacl_a_00527

[5] Dhruv, Kumar., Robin, Cohen., Lukasz, Golab. (2019). Online abuse detection: the value of preprocessing and neural attention models.. doi: 10.18653/V1/W19-1303

[6] Rui, Song., Fausto, Giunchiglia., Qiang, Shen., Nanjun, Li., Hao, Xu. (2022). Improving Abusive Language Detection with Online Interaction Network. Information Processing and Management, doi 10.1016/j.ipm.2022.103009

[7] Kunze, Wang., Dong, Lu., Caren, Han., Siqu, Long., Josiah, Poon. (2020). Detect All Abuse! Toward Universal Abusive Language Detection Models. doi: 10.18653/V1/2020.COLING-MAIN.560

[8] Dhruv, Kumar., Robin, Cohen., Lukasz, Golab. (2019). Online abuse detection: the value of preprocessing and neural attention models. doi: 10.18653/V1/W19-1303

[9] Santhosh, Rajamanickam., Pushkar, Mishra., Helen, Yannakoudakis., Ekaterina, Shutova. (2020). Joint Modelling of Emotion and Abusive Language Detection. doi: 10.18653/V1/2020.ACL-MAIN.394

[10] S. Joshi and B. Nigam, "Categorizing the Document Using Multi Class Classification in Data Mining," 2011 International Conference on Computational Intelligence and Communication Networks, Gwalior, India, 2011, pp. 251-255, doi: 10.1109/CICN.2011.50.

[11] Lokesh, Rajulapati., Sivadurgaprasad, Chinta., Bala, Shyamala., Raghunathan, Rengaswamy. (2022). Integration of Machine Learning and First Principles Models. Aiche Journal, doi: 10.1002/aic.17715

[12] Yuxin, Pei. (2023). A Comparative Study of Machine Learning and Automatic Machine Learning Models for Facial Mask Recognition. doi: 10.1109/ICCCS57501.2023.10151333

[13] B. Nigam, A. Nigam, P. Dalal, "Comparative Study of Top 10 Algorithms for Association Rule Mining," *International Journal of Computer Sciences and Engineering*, Vol. 5, Issue 8, pp. 190-195, August 2017. E-ISSN: 2347-2693.
[14] Maximilian, Pichler. (2023). Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution, doi: 10.1111/2041-210x.14061
[15] B. Nigam, P. Ahirwal, S. Salve, S. Vamney, "Document Classification Using Expectation Maximization with Semi-Supervised Learning," *International Journal on Soft Computing (IJSC)*, Vol. 2, No. 4, pp. 37-44, November 2011. DOI: 10.5121/ijsc.2011.2404.
[16] Shams, Forruque, Ahmed., Md., Sakib, Bin, Alam., M.Alaa, Ahmed, Hassan., Mahtabin, Rodela, Rozbu., Taoseef, Ishtiak., Nazifa, Rafa., M., Mofijur., A., B., M., Shawkat, Ali., Amir, H., Gandomi. (2023). Deep learning modelling techniques: progress, applications, advantages, and challenges. Artificial Intelligence Review, doi: 10.1007/s10462-023-10466-8
[17] R. Alharthi, R. Alharthi, R. Shekhar and A. Zubiaga, "Target-Oriented Investigation of Online Abusive Attacks: A Dataset and Analysis," in IEEE Access, vol. 11, pp. 64114-64127, 2023, doi: 10.1109/ACCESS.2023.3289148.
[18] Hoyun, Song., Soo, Hyun, Ryu., Huije, Lee., Jong, Park. (2021). A Large-scale Comprehensive Abusiveness Detection Dataset with Multifaceted Labels from Reddit.
[19] Noman, Ashraf., Arkaitz, Zubiaga., Alexander, Gelbukh. (2021). Abusive language detection in YouTube comments leveraging replies as conversational context. PeerJ, doi: 10.7717/PEERJ-CS.742
[20] Bertie, Vidgen., Dong, Nguyen., Helen, Margetts., Patrícia, Rossini., Rebekah, Tromble. (2021). Introducing CAD: the Contextual Abuse Dataset. doi: 10.18653/V1/2021.NAACL-MAIN.182