IJO - INTERNATIONAL JOURNAL OF EDUCATIONAL RESEARCH
(E.ISSN: 2805-413X) (P.ISSN: 2536-6696) Mullah Nanlir Sallau *
https://ijojournals.com/ Volume 08 || Issue 03 || March, 2025 ||
"Offensive Text Detection using Deep Learning: A Review with Open Challenges"

Offensive Text Detection using Deep Learning: A Review with Open Challenges

Mullah Nanlir Sallau, Ladan Nanbal Jibba, Ramson Emmanuel Nannim, Gokir Justine Ali, Emmanuel Datti Useni, Dashe Miapmuk Obadia

Department of computer Science Education,
Federal University of Education Pankshin, Plateau State, Nigeria.

Abstract

The rapid expansion of online communication platforms has led to an increase in offensive and harmful content, necessitating robust detection mechanisms. Deep learning techniques have emerged as a powerful approach for offensive text detection, leveraging neural networks to capture complex linguistic patterns and contextual nuances. This paper provides a comprehensive review of deep learning-based methods for offensive text detection, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), gated recurrent units (GRU), and transformer-based models such as BERT and GPT. We discuss key challenges such as class imbalance, context understanding, bias mitigation, and the interpretability of deep learning models. Furthermore, we explore evaluation metrics, benchmark datasets, and recent advancements in adversarial robustness and explainability. Finally, we highlight future research directions to improve the effectiveness, fairness, and scalability of deep learning techniques in offensive text detection.

KeyWords: Deep learning, online, Text, Neural networks

Introduction

The exponential growth of digital communication platforms has revolutionised the way individuals interact, enabling seamless global connectivity. However, this proliferation has also led to the widespread dissemination of offensive and harmful content, posing significant challenges for platform moderators and policymakers(Ranasinghe & Zampieri, 2021). Offensive language, including hate speech, cyberbullying, and abusive comments, can have severe social and psychological consequences, necessitating the development of robust detection mechanisms(Mehta, 2024). While traditional approaches such as rule-based and lexicon-based filtering have been employed for text moderation, they often fall short in handling the complex, context-dependent, and evolving nature of offensive language(Mullah & Zainon, 2021). As a result, deep learning techniques have gained significant traction in offensive text detection, offering enhanced capabilities in recognising subtle and implicit forms of harmful speech.

Deep learning models leverage large-scale datasets to learn intricate patterns in textual data, enabling automated detection systems to differentiate offensive content from benign expressions. Various architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), have demonstrated remarkable success in capturing semantic and syntactic nuances. Unlike traditional keyword-based methods, deep learning models can analyse text holistically, considering contextual cues, sentiment, and linguistic variations to enhance classification accuracy. The effectiveness of these models is further amplified by advancements in natural language processing (NLP), particularly in pre-trained language models and transfer learning(Mullah & Zainon, 2022; Sun et al., 2019).

Despite their successes, deep learning-based approaches for offensive text detection encounter several challenges. One of the primary concerns is the inherent bias present in training datasets, which can lead to skewed predictions and ethical concerns. Additionally, offensive language is highly dynamic, evolving over time with new slangs(Dadvar & Eckert, 2020), coded words(Gamal et al., 2023), and adversarial attacks(Zhang et al., 2020) designed to bypass detection mechanisms. Another challenge is the computational complexity associated with deploying large-scale deep learning models, which may hinder real-time detection in resource-constrained environments. Addressing these issues requires continuous research in data augmentation, bias mitigation, model interpretability, and adversarial robustness.

This review aims to provide a comprehensive analysis of the state-of-the-art deep learning techniques for offensive text detection, evaluating their strengths, limitations, and potential advancements. By synthesising existing literature, this paper highlights key research gaps and outlines future directions to improve the efficiency and fairness of automated detection systems.

Motivation

The increasing prevalence of offensive language on digital platforms has raised serious concerns regarding online safety, mental health, and ethical responsibilities. Social media networks, comment sections, and online forums have become breeding grounds for hate speech, cyber harassment, and discrimination, often leading to severe consequences for individuals and communities. Traditional moderation methods, such as manual review and rule-based filtering, are inadequate in handling the scale and complexity of modern online discourse. The need for more sophisticated, automated solutions that can detect and mitigate offensive content in real-time has become paramount.

Deep learning has emerged as a powerful tool to address these challenges, providing robust frameworks that can analyse vast amounts of textual data with high accuracy. By leveraging machine learning techniques, platforms can enhance their content moderation strategies, reduce human workload, and ensure a safer digital environment. Moreover, as offensive language continues to evolve, the adaptability of deep learning models allows them to keep up with emerging trends and linguistic variations. The motivation behind this review is to explore and assess the latest advancements in deep learning-based offensive text detection, identify gaps in existing research, and propose future directions to improve the reliability, fairness, and efficiency of these systems.

Objectives of the Review

This review aims to achieve the following objectives:

1. To analyse and compare various deep learning techniques employed in offensive text detection, including CNNs, RNNs, LSTMs, and transformer-based models.

2. To identify the limitations and challenges associated with deep learning-based offensive text detection, focusing on bias, adversarial robustness, and computational efficiency.

3. To explore potential future directions and advancements in offensive text detection, including bias mitigation strategies, real-time processing improvements, and cross-lingual approaches.

By addressing these objectives, this review seeks to contribute to the ongoing development of more effective, fair, and scalable automated moderation systems for detecting offensive content online.

Literature Review

The literature on offensive text detection using deep learning techniques has expanded significantly in recent years(Wan et al., 2024; Yang et al., 2023). Various studies have explored different methodologies, datasets, and architectures to enhance the accuracy and efficiency of automated detection systems. This section provides a detailed review of existing research, categorising studies based on their approaches, challenges, and contributions.

Traditional Approaches to Offensive Text Detection

Early offensive text detection systems relied on rule-based and lexicon-based approaches, where predefined keywords and phrases were used to filter harmful content(Pradhan et al., 2020). While these methods provided a foundational framework, they were highly limited in detecting nuanced and context-dependent offensive language. Machine learning-based models, such as support vector machines (SVM) and logistic regression, later emerged, leveraging handcrafted features like n-grams, sentiment analysis, and syntactic patterns(Hemmatian & Sohrabi, 2019).

Deep Learning-Based Approaches

With advancements in natural language processing, deep learning models have become the dominant approach for offensive text detection. Key architectures include(Li et al., 2020):

1. Convolutional Neural Networks (CNNs): CNNs have been widely used for text classification due to their ability to capture local dependencies in textual data. Studies have shown that CNNs perform well in detecting explicit offensive content but struggle with contextual understanding(Sachdeva et al., 2021).

2. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks: RNNs and LSTMs are effective in capturing sequential dependencies in text, making them suitable for detecting implicit offensive language and sarcasm. However, they suffer from long-term dependency issues and computational inefficiency(Fati et al., 2023).

3. Transformer-Based Models: The introduction of transformer models, particularly BERT and its variants, has revolutionised offensive text detection. These models leverage attention mechanisms to understand context more effectively, outperforming traditional deep learning approaches in various benchmarks(Zaheer et al., 2020).

Table 1: Comparison of ML, DL and LLM

Feature	**Machine Learning (ML)**	**Deep Learning (DL)**	**Large Language Models (LLM)**
Approach	Uses feature engineering with traditional models (e.g., SVM, Naïve Bayes)	Uses neural networks like CNNs, RNNs, LSTMs, Transformers	Uses large-scale pre-trained models like GPT, BERT
Data Requirement	Requires structured datasets with hand-crafted features	Requires large labeled datasets	Requires massive pre-trained datasets, but fine-tuning needs labeled data
Feature Engineering	Necessary (e.g., TF-IDF, n-grams)	Minimal (automated feature extraction)	Not needed (learns from context)
Computational Cost	Low to moderate	High	Very high (needs GPUs/TPUs)
Training Time	Fast (minutes to hours)	Moderate (hours to days)	Very long (days to weeks)
Accuracy	Moderate	High	Very high
Interpretability	High (clear rules, explainable features)	Medium (some explainability techniques available)	Low (black-box nature)
Handling Context	Limited (bag-of-words, basic n-grams)	Better (sequential dependencies)	Excellent (understands deep context, sarcasm, idioms)
Scalability	Scales well for small to medium datasets	Scales well but needs specialized hardware	Difficult to scale without significant infrastructure
Real-time Performance	Fast (can run on CPU)	Moderate (needs GPU for best performance)	Slower (requires large compute resources)
Pre-trained Models	Usually trained from scratch	Some pre-trained embeddings (e.g., word2vec, GloVe)	Uses state-of-the-art pre-trained models (BERT, GPT, etc.)
Example Models	SVM, Naïve Bayes, Random Forest	CNN, RNN, LSTM, Transformer	GPT-4, BERT, T5
Best Use Case	Small datasets, explainability required	Large datasets, complex text understanding needed	High accuracy, nuanced offensive text detection, detecting sarcasm or implicit bias

Machine Learning (ML) is good for small datasets with explainability needs. Deep Learning (DL) provides better accuracy but needs more data and computational resources. Large Language Models (LLM) offer state-of-the-art performance but are computationally expensive and harder to interpret.

Table 2: The Advantages and limitations of CNN, RNNLSTM and BERT

Model	**Advantages**	**Limitations**
CNNs (Convolutional Neural Networks)	- Good at capturing local n-gram features in text.	CNNs (Convolutional Neural Networks)
RNNs (Recurrent Neural Networks)	- Can capture sequential dependencies in text.	RNNs (Recurrent Neural Networks)
LSTMs (Long Short-Term Memory Networks)	- Better at handling long-range dependencies compared to RNNs.	STMs (Long Short-Term Memory Networks)
BERT-Based Models (e.g., BERT, RoBERTa, DistilBERT)	- Superior contextual understanding with bidirectional attention.	BERT-Based Models (e.g., BERT, RoBERTa, DistilBERT)

Evaluation Metrics for Offensive Text Detection

Evaluating offensive text detection models requires various classification metrics to measure accuracy, fairness, and robustness. In classification tasks, especially in offensive text detection, predictions can be categorized into four types based on the actual and predicted labels. Here are the key metrics(Atalan Ergin et al., 2021; Peck et al., 2024; Wan et al., 2024):

1. True Positive (TP)
Definition: The model correctly predicts an instance as offensive when it is actually offensive.
Example:
Actual: "You are an idiot!" → Offensive
Predicted: Offensive (Correct)
Importance: High TP ensures that offensive content is successfully identified.

2. True Negative (TN)
Definition: The model correctly predicts an instance as non-offensive when it is actually non-offensive.
Example:
Actual: "Hope you have a great day!" → Non-Offensive
Predicted: Non-Offensive (Correct)
Importance: High TN ensures that normal text is not flagged incorrectly.

3. False Positive (FP) (Type I Error)
Definition: The model incorrectly predicts an instance as offensive when it is actually non-offensive.
Example:
Actual: "You're so crazy! 😂" → Non-Offensive
Predicted: Offensive (Incorrect)
Issue: Can lead to over-censorship, where harmless content is wrongly removed.

4. False Negative (FN) (Type II Error)
Definition: The model incorrectly predicts an instance as non-offensive when it is actually offensive.
Example:

Actual: "I hate your existence!" → Offensive
Predicted: Non-Offensive (Incorrect)
Issue: Can allow harmful or abusive content to go undetected, which is dangerous in social media moderation.

Why These Terms Matter?
High TP & TN = Good model performance.
High FP = Too many false alarms (overblocking).
High FN = Misses harmful content (unsafe environment).

Based on the research objectives, BERT-based models are the most suitable for offensive text detection due to the following reasons:

Comprehensive Context Understanding (Objective 1)

Unlike CNNs, RNNs, and LSTMs, BERT utilizes a bidirectional transformer mechanism, meaning it considers both left and right contexts simultaneously.This is crucial for detecting offensive content where the meaning depends on surrounding words (e.g., sarcasm, implicit hate speech).

Addressing Bias and Robustness (Objective 2)

BERT-based models can be fine-tuned with debiasing techniques (e.g., adversarial training, fairness-aware learning).Their ability to generalize across different text patterns makes them more resilient to adversarial attacks compared to RNNs and CNNs.

Scalability and Future Advancements (Objective 3)

Transformer models, especially lightweight variants like DistilBERT and ALBERT, allow for improved real-time processing.Multilingual BERT (mBERT) and XLM-R support cross-lingual offensive text detection, aligning with future research directions.Further research on low-resource fine-tuning and efficient transformer variants can enhance computational efficiency.

Table 3: Pros and Cons of evaluation metrics

Metric	**Formula**	**Description**	**Pros**	**Cons**
Accuracy	Ac = (TP + TN) / (TP + TN + FP + FN)	Measures overall correctness	Simple and intuitive	Not useful for imbalanced datasets
Precision (Positive Predictive Value)	Pr = TP / (TP + FP)	Measures correctness of offensive text predictions	Reduces false positives	Can ignore false negatives
Recall (Sensitivity)	Rc = TP / (TP + FN)	Measures how well the model detects	Reduces false negatives	Can increase false positives

		offensive text		
F1-Score	F1 = 2 × (Pr × Rc) / (Pr + Rc)	Balances precision and recall	Works well for imbalanced data	Doesn't show absolute errors
ROC-AUC (Area Under the Curve)	Computed from the ROC curve	Measures model performance across thresholds	Works well for balanced data	May be misleading for imbalanced data

Open Challenges

The following are the open challenges in offensive text detection that still require more attention for holisticsolution this problem. Tenof key open challenges in offensive text detection are expatiated below for better understanding and handling:

1. Bias in Training Data: Training datasets often contain societal biases, leading to unfair classifications that disproportionately impact certain demographic groups. Addressing these biases requires dataset balancing, fairness-aware algorithms, and adversarial debiasing techniques.

2. Contextual Understanding: Sarcasm, metaphors, and indirect offensive expressions require deeper contextual comprehension, which many models struggle to capture. Improved NLP techniques and external knowledge integration are needed to enhance contextual awareness.

3. Adversarial Robustness: Offensive users intentionally manipulate text using obfuscation techniques (e.g., leetspeak, misspellings). Developing robust models that can recognize these alterations is essential to maintain detection accuracy.

4. Scalability: Deploying deep learning models at scale is computationally expensive. Efficient model architectures, quantization, and pruning techniques are required for real-time processing on large-scale platforms.

5. Multilingual Detection: Offensive language varies across cultures and languages. Expanding datasets and developing cross-lingual transfer learning approaches are crucial for accurate multilingual detection.

6. Explainability: Many deep learning models function as black boxes, making it difficult to interpret their decisions. Enhancing model transparency through explainable AI techniques is necessary for accountability.

7. Data Privacy: Protecting user data while training detection models is a significant concern. Techniques like federated learning and differential privacy can help address privacy risks.

8. Ethical Considerations: Ensuring a balance between free speech and content moderation requires ethical AI guidelines to avoid unjust censorship.

9. Multimodal Analysis: Offensive content is not limited to text but often appears in multimedia formats. Developing models that integrate text, image, and video analysis is a growing necessity.

10. Continuous Learning: Offensive language evolves rapidly, necessitating models that can adapt over time. Continuous learning frameworks and reinforcement learning can help improve adaptability.

Conclusion

Offensive text detection is a crucial challenge in the digital era, where harmful content can have severe social, psychological, and legal consequences. Deep learning techniques have significantly advanced the field by enabling models to capture intricate linguistic patterns, contextual meanings, and implicit biases in textual data. In this review, we explored various deep learning architectures, including CNNs, RNNs, LSTMs, GRUs, and state-of-the-art transformer-based models such as BERT and GPT, highlighting their strengths and limitations in offensive text detection.

While deep learning models exhibit superior performance compared to traditional machine learning methods, they also present notable challenges. Issues such as dataset imbalance, model interpretability, and unintended biases remain major obstacles. Moreover, offensive language is highly context-dependent, requiring models to distinguish between genuinely harmful content and benign expressions such as sarcasm or humor. Addressing these challenges requires ongoing research in explainability, adversarial robustness, and fairness in AI.

Evaluation metrics play a vital role in assessing model performance, with precision, recall, F1-score, and advanced measures like Matthews Correlation Coefficient (MCC) and PR-AUC helping to provide a comprehensive understanding of a model’s effectiveness. However, real-world deployment demands more than just high accuracy; fairness, scalability, and real-time performance must also be prioritized to ensure practical usability in social media moderation, content filtering, and automated monitoring systems.

Future research should focus on refining model architectures to improve contextual understanding and reduce false positives and false negatives. Additionally, integrating multimodal learning, where text is analysed alongside images, videos, or metadata, could enhance detection capabilities. The development of ethically responsible AI, incorporating fairness-aware algorithms and bias-mitigation strategies, is also crucial for ensuring inclusive and unbiased offensive text detection systems.

In conclusion, deep learning has revolutionized offensive text detection, but continuous advancements are needed to build more robust, fair, and scalable models. By addressing current limitations and leveraging interdisciplinary research in linguistics, ethics, and AI, the field can progress toward safer and more responsible digital communication environments.

Acknowledgement:

This work has been supported by Tetfund IBR grant in Nigeria.

References

Atalan Ergin, D., Akgül, G., & Güney Karaman, N. (2021). Ethnic-based cyberbullying: The role of adolescents' and their peers' attitudes towards immigrants. *Turkish Journal of EducationTURJE 2021*, *10*(2). https://doi.org/10.19128/turje.879347

Dadvar, M., & Eckert, K. (2020). Cyberbullying detection in social networks using deep learning based models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12393 LNCS*, 245–255. https://doi.org/10.1007/978-3-030-59065-9_20

Fati, S. M., Muneer, A., Alwadain, A., & Balogun, A. O. (2023). Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction. *Mathematics*, *11*(16). https://doi.org/10.3390/math11163567

Gamal, D., Alfonse, M., Jiménez-Zafra, S. M., & Aref, M. (2023). Intelligent Multi-Lingual Cyber-Hate Detection in Online Social Networks: Taxonomy, Approaches, Datasets, and Open Challenges. *Big Data and Cognitive Computing*, *7*(2). https://doi.org/10.3390/bdcc7020058

Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, *52*(3), 1495–1545. https://doi.org/10.1007/s10462-017-9599-6

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). *A Survey on Text Classification: From Shallow to Deep Learning*. *September*. http://arxiv.org/abs/2008.00364

Mehta, K. (2024). Impact Of Cyber Bullying In Teenagers. *Multi-Disciplinary Journal*, *1*(2), 1–14. www.mahratta.org,editor@mahratta.org

Mullah, N. S., & Zainon, W. M. N. W. (2022). Improving detection accuracy of politically motivated cyber-hate using heterogeneous stacked ensemble (HSE) approach. *Journal of Ambient Intelligence and Humanized Computing*, 1–12. https://doi.org/https://doi.org/10.1007/s12652-022-03763-7

Mullah, N. S., & Zainon, W. M. Na. W. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media : A Review. *IEEE Access*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Peck, H. C., Tzani, C., Lester, D., Williams, T. J. V., & Page, J. (2024). Cyberbullying in the UK: The Effect of Global Crises on the Victimization Rates. *Journal of School Violence*, *23*(1), 111–123. https://doi.org/10.1080/15388220.2023.2278473

Pradhan, R., Chaturvedi, A., Tripathi, A., & Sharma, D. K. (2020). A review on offensive language detection. *Lecture Notes in Networks and Systems*, *94*(January), 433–439. https://doi.org/10.1007/978-981-15-0694-9_41

Ranasinghe, T., & Zampieri, M. (2021). *Multilingual Offensive Language Identification for Low-resource Languages*. http://arxiv.org/abs/2105.05996

Sachdeva, J., Chaudhary, K. K., Madaan, H., & Meel, P. (2021). Text Based Hate-Speech Analysis. *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, *January*, 661–668. https://doi.org/10.1109/ICAIS50930.2021.9396013

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11856 LNAI*(2), 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

Wan, Y., Bi, Z., He, Y., Zhang, J., Zhang, H., Sui, Y., Xu, G., Jin, H., & Yu, P. (2024). Deep Learning for Code Intelligence: Survey, Benchmark and Toolkit. *ACM Computing Surveys*. https://doi.org/10.1145/3664597

Yang, X., Song, Z., King, I., & Xu, Z. (2023). A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 8934–8954. https://doi.org/10.1109/TKDE.2022.3220219

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, *2020-Decem*(NeurIPS).

Zhang, W. E. I. E., Sheng, Q. Z., & Alhazmi, A. (2020). Adversarial Attacks on Deep Learning Models in Natural Language Processing : A Survey. *ACM Transactions on Intelligent Systems and Technology*, *1*(1).