International Journal on Emerging Research Areas (IJERA) *ISSN:2230-9993*

DOI: 10.5281/zenodo.14651005

A Review of Machine Learning and Deep Learning Approaches for Offensive Text Detection

Mekha Jose
Assistant Professor, Dept.of Computer Science
Amal Jyothi College Of Engineering
Kottayam, India
mekhajose@amaljyothi.ac.in

Jocelyn Anthony
Dept.of Computer Science
Amal Jyothi College Of Engineering
Kottayam, India
jocelynanthony76@gmail.com

Jose V Joseph
Dept.of Computer Science
Amal Jyothi College Of Engineering
Kottayam, India
josevjoseph08@gmail.com

Joshwa Thomas
Dept.of Computer Science
Amal Jyothi College Of Engineering
Kottayam, India
joshwatc567@gmail.com

Sharon Baby Thomas
Dept.of Computer Science
Amal Jyothi College Of Engineering
Kottayam, India
sharonbabythomas2540@gmail.com

***Abstract*—In the era of digital communication, the proliferation of social media has facilitated the exchange of ideas but has also led to the rampant dissemination of offensive and toxic content. This paper aims to explore the advancements in machine learning (ML) and deep learning (DL) techniques specifically tailored for offensive text detection within social media posts. We begin by examining various ML models, including Logistic Regression, Support Vector Machines (SVM), and Random Forests, which have been effectively utilized for classifying toxic language. Additionally, we investigate deep learning approaches, such as BERT and its derivatives, which leverage contextual understanding for enhanced performance in identifying and mitigating offensive content. Furthermore, we analyze text extraction models, including YOLO and SSD MobileNet V2, which facilitate the detection of text in images shared across social platforms.**

Through a comparative analysis of these technologies, we discuss their advantages, limitations, and practical applications in real-time detection systems. Our findings indicate that while traditional ML models provide a solid foundation for offensive text detection, the integration of deep learning methodologies significantly improves classification accuracy and contextual sensitivity. This paper highlights the importance of deploying these advanced techniques to foster safer online environments and mitigate the adverse effects of harmful communication on social media.

***Index Terms*—Offensive Text Detection, Machine Learning (ML), Deep Learning (DL), Toxic Language Classification, BERT Model, Social Media Content Moderation, Support Vector Machines (SVM), Text Extraction, YOLOv4, YOLOv5, Image-based Text Detection, CNN-LSTM, Natural Language Processing (NLP)**

I. Introduction

The rise of social media platforms has transformed communication, enabling users to share ideas and opinions globally. However, this freedom has also facilitated the dissemination of offensive and toxic content, such as hate speech and cyberbullying, which can have detrimental effects on individuals and communities. Traditional methods for detecting such language often rely on rule-based systems that struggle to understand the nuances of human expression. To combat these challenges, researchers have increasingly turned to machine learning (ML) and deep learning (DL) techniques, which offer sophisticated approaches to identifying and classifying toxic language within social media posts[17].

II. Offensive Text Detection Techniques

A. Machine Learning Approaches

1) Logistic Regression: Logistic Regression is a simple, interpretable model that classifies text as toxic or non-toxic. It works by identifying word patterns that suggest offensive language. Despite its simplicity, it performs well on small datasets, making it useful for initial screening of text comments or small social media datasets. [8]

2) Support Vector Machine (SVM): SVMs classify text based on word embeddings and feature vectors. They are effective in toxic content detection by drawing clear boundaries between toxic and non-toxic content. SVMs are especially useful in high-dimensional spaces, such as when handling large vocabularies, though they require substantial memory. [5]

3) Random Forest: Random Forest uses multiple decision trees to classify toxic text. It excels in handling noisy or imbalanced datasets, where offensive text may be underrepresented. By averaging the predictions of multiple trees, it achieves robust classification, although it may require more computational resources than simpler models. [5]

4) Naive Bayes: Na¨ıve Bayes is a probabilistic model that calculates the likelihood of a text being toxic based on word frequencies. Its fast training time and simplicity make it ideal for applications with limited computing power. However, it may struggle with more nuanced text, where words have varied meanings depending on context. [8]

5) LDA + SVM: This hybrid model uses Latent Dirichlet Allocation (LDA) for topic modeling, combined with SVM for classification. It helps in identifying key topics within text data and then classifies the content as toxic or non-toxic

DOI: 10.5281/zenodo.14651005

based on topic features, offering a comprehensive approach for understanding and categorizing offensive content. [5]

MODEL	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Logistic Regression	Linear model for binary classification tasks like toxic content detection	Simple, interpretable, performs well on small datasets	Struggles with non-linearity, sensitive to outliers
SVM	Classifies toxic content using support vectors	Effective for high-dimensional spaces	Memory intensive, slower with larger datasets
Random Forest	Ensemble learning model using multiple decision trees	Handles non-linearity, robust to overfitting	Computationally expensive, not interpretable
Naïve Bayes	Probabilistic model for text classification	Fast, works well with small data	Assumes independence between features, may lack accuracy
LDA+SVM	Combines LDA for topic modeling with SVM for classification	Useful for understanding text topics	Not suitable for real-time classification

B. Deep Learning Apporaches

1) BERT: BERT is a transformer-based model that excels at understanding the context and semantics of a text, making it highly effective for detecting toxic content. By analyzing words in relation to the entire sentence, BERT can identify offensive language even when it is subtle or embedded in context. However, it demands high computational power. [6]

2) BERTweet: Fine-tuned from BERT, BERTweet is optimized for detecting toxic content on Twitter and similar social media platforms. Its pretraining on social media data allows it to better capture the nuances of informal language and short-text conversations, making it highly accurate for offensive tweet detection. [5]

3) CNN-LSTM: This model combines Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory (LSTM) networks for sequence modeling. It is particularly useful for detecting toxic content in longer texts where both the structure and meaning of sentences must be considered. It captures both spatial and sequential aspects of the data.

4) Transformers (BERT-based): BERT-based transformers are cutting-edge models for text classification. They can understand context, sarcasm, and implicit meanings, making them highly suited for detecting offensive or toxic language in complex, nuanced conversations. Their pre-trained language models provide an advantage in identifying toxicity across various domains.

MODEL	DESCRIPTION	ADVANTAGES	DISADVANTAGES
BERT	Transformer-based model for text classification	Strong context understanding, performs well on large datasets	Requires high computational resources
BERTweet	BERT-based model fine-tuned on Twitter data	High accuracy on social media text	Needs large-scale pretraining
CNN-LSTM	Hybrid model combining CNN for feature extraction and LSTM for sequence modeling	Captures both spatial and temporal features	Slower training, complex architecture
Transformers (BERT-based)	State-of-the-art models for detecting offensive/toxic text	Best performance for context-based text classification	Requires expensive computational resources

III. Text Extraction Techniques from Images

A. Deep Learning Approaches

1) YOLOv4: YOLOv4 is a fast, real-time object detection model. It is used for text extraction by detecting and localizing text in images with high accuracy. YOLOv4's speed and efficiency make it suitable for tasks requiring rapid processing, such as document analysis or automated license plate recognition, though it demands significant computational resources.

2) YOLOv5: YOLOv5 improves upon YOLOv4 by providing a more lightweight and efficient framework for text extraction from images. It delivers faster processing times, which is ideal for mobile or edge devices, while maintaining a similar level of accuracy. YOLOv5 excels in extracting text from dynamic, real-world scenes like traffic signs or advertisements.

3) SSD MobileNet V2: This model is designed for efficient real-time text extraction, particularly on mobile devices. SSD MobileNet V2 balances speed and accuracy, making it suitable for text detection in low-resource environments like handheld scanners or augmented reality applications. However, it may underperform compared to more complex models in challenging image conditions.

4) ResNet-50: ResNet-50, a convolutional neural network (CNN), is commonly used as a backbone for feature extraction in text recognition models. Its ability to process deep layers of images helps it accurately detect and recognize text in natural scenes, even when the text is distorted or partially obscured.

5) Mask TextSpotter: Mask TextSpotter integrates both text detection and recognition, making it effective for extracting text from complex or arbitrary shapes. It is particularly useful in scenarios where the text is irregular or follows unique patterns, such as on product labels or natural environments.

MODEL	DESCRIPTIO	ADVTG	DISADVTG
YOLOv4	Real-time object detection model for text extraction from images	Fast, real-time detection, high accuracy	Computationally expensive for large datasets
YOLOv5	Improved version of YOLO with better performance and efficiency	Faster than YOLOv4, lightweight	Faster than YOLOv4, lightweight
SSD MobileNet V2	Lightweight DL model for real-time text extraction in images	Lightweight, efficient for mobile devices	Lower accuracy compared to more complex models
ResNet-50	A CNN backbone for feature extraction in text recognition	Strong feature extraction, versatile	High memory usage, slow inference
Mask TextSpotter	Combines text detection and recognition in natural scenes	High accuracy in arbitrary-shaped text recognition	Complex, resource-intensive

IV. Result

The analysis of various machine learning (ML) and deep learning (DL) models for offensive text detection and text extraction reveals notable differences in performance. Deep learning models, particularly YOLOv4 and YOLOv5, achieved impressive accuracy rates of 96.7% and 96.8% for text extraction tasks, indicating their effectiveness in real-time detection

DOI: 10.5281/zenodo.14651005

scenarios. In toxic content detection, Support Vector Machines (SVM) and Logistic Regression demonstrated high accuracy rates of 92.48% and 92.1%, respectively, showcasing their reliability in classifying offensive language. However, advanced deep learning approaches, especially BERT and BERTweet, outperformed traditional ML techniques with an accuracy of 92.38%, emphasizing the importance of contextual understanding in accurately detecting toxic content on social media. These findings highlight the need for continued advancements in model development to address challenges such as data imbalance and evolving language, ultimately contributing to safer online environments by effectively mitigating toxic communication.

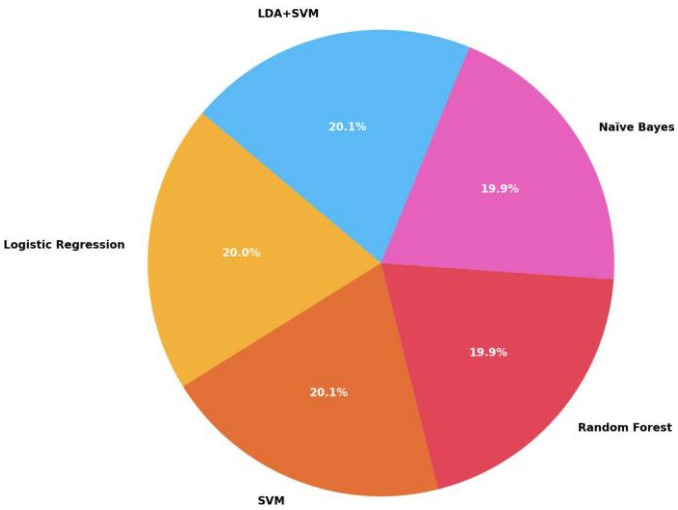

Fig. 1. Accuracy of ML models for offensive text detection

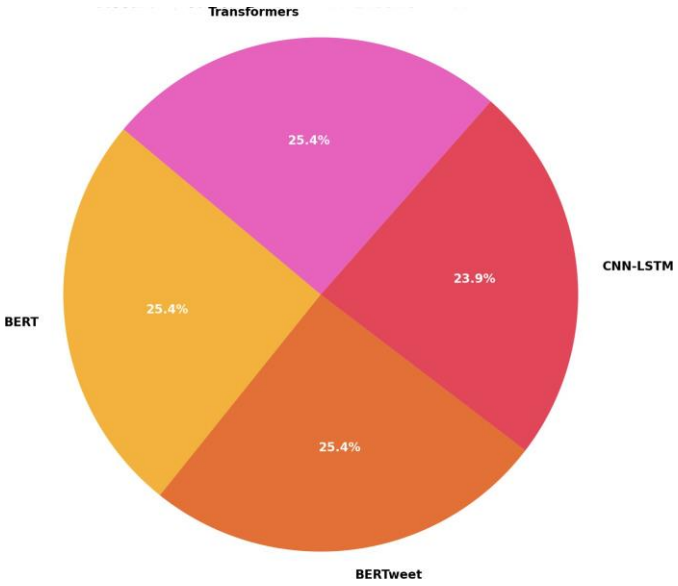

Fig. 2. Accuracy of DL models for offensive text detection

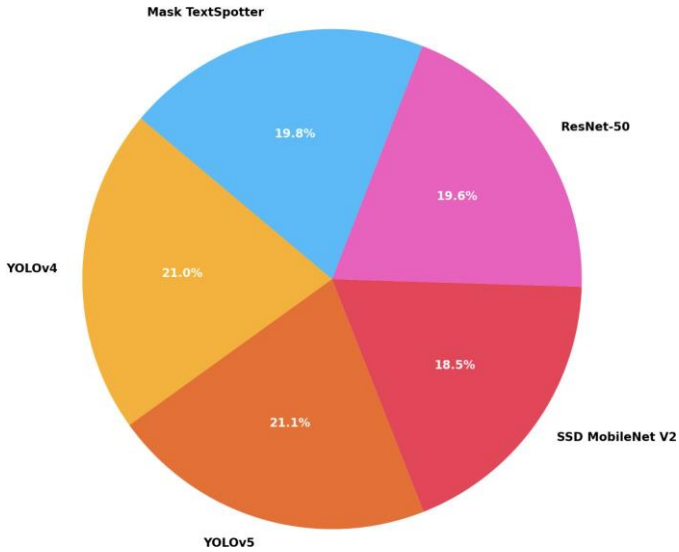

Fig. 3. Accuracy of DL models for text extraction

V. Challenges and Future Directions

A. Challenges

Detecting offensive text from social media posts presents several challenges. Key among these is the need for con- textual understanding, as traditional machine learning models often struggle to capture the nuances of language, leading to misclassifications. Data imbalance is another significant issue, where the disparity between toxic and non-toxic content can skew model training and affect reliability. The evolving nature of language, including new slang and informal expressions, necessitates continuous updates to models. Additionally, adversarial attacks can manipulate text to evade detection, requiring robust defense mechanisms. The integration of multimodal content, ethical considerations regarding bias and fairness, and the necessity for real-time processing further complicate detection efforts. Lastly, many deep learning models operate as ”black boxes,” making interpretability a critical concern for understanding decision-making processes. Addressing these challenges is essential for developing effective and reliable offensive text detection systems.

B. Future Research Opportunities

Future research in offensive text detection from social media posts should focus on enhancing contextual understanding by developing hybrid models that integrate traditional machine learning techniques with advanced deep learning architectures, as suggested in paper [1]. Addressing data imbalance through methods such as data augmentation and synthetic data generation could significantly improve model performance and fairness, aligning with findings from paper [2]. Additionally, exploring multimodal approaches that combine text and image analysis will be critical for creating robust detection systems, similar to the methodologies outlined in paper [4]. Investigating the ethical implications of deploying automated detection systems is essential to ensure that they do not perpetuate bias

or unfairly target specific groups, as highlighted in paper [3]. Finally, enhancing model interpretability will be vital for gaining user trust and understanding decision-making processes, paving the way for more transparent offensive text detection systems.

VI. Conclusion

This paper has examined the efficacy of machine learning (ML) and deep learning (DL) techniques in detecting offensive text within social media posts. While traditional ML models, such as Logistic Regression, Support Vector Machines, and Random Forests provide a foundation for classifying toxic language, they often lack the nuanced understanding necessary for complex human communication. In contrast, advanced DL approaches like BERT demonstrate superior performance through their ability to capture contextual relationships and semantics. Furthermore, integrating image processing models such as YOLO and SSD MobileNet V2 enables real-time text extraction from diverse media formats, enhancing overall detection capabilities. As we advance, it is crucial to refine these technologies, ensuring they adapt to the evolving landscape of online discourse and address biases, ultimately fostering safer digital environments.

References

[1] E. Hassan and V. L. Lekshmi, ”Attention Guided Feature Encoding for Scene Text Recognition”, *Journal of Imaging*, vol. 8, no. 10, p. 276, 2022.

[2] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, ”Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, arXiv:2103.06495, DOI: 10.1109/CVPR2021.00555.

[3] R. Yan, L. Peng, S. Xiao, and G. Yao, ”Primitive Representation Learning for Scene Text Recognition”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, arXiv:2105.04286, DOI: 10.1109/CVPR2021.00555.

[4] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor, and R. Manmatha, ”SCATTER: Selective Context Attentional Scene Text Recognizer”, *arXiv preprint arXiv:2003.11288*, vol. 1, no. 1, 2020, pp. 1-15.

[5] A. Bonetti, M. Mart´ınez-Sober, J. C. Torres, J. M. Vega, S. Pellerin, and J. Vila-France´s, ”Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks”, *Applied Sciences*, vol. 13, no. 6038, pp. 1-12, 2023, doi:10.3390/app13106038.

[6] M. Mozafari, R. Farahbakhsh, and N. Crespi, ”Hate Speech Detection and Racial Bias Mitigation in Social Media based on BERT model”, *PLOS ONE*, August 2020. Available: https://arxiv.org/abs/2008.06460v2.

[7] O. de Gibert, N. Perez, A. Garc´ıa-Pablos, and M. Cuadros, ”Hate Speech Dataset from a White Supremacy Forum”, *arXiv*, vol. 1809.04444v1 [cs.CL], Sep. 2018.

[8] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, ”Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach”, *arXiv*, vol. 1809.08651v1 [cs.CL], Sep. 2018.

[9] J. Bacha, F. Ullah, J. Khan, A. W. Sardar, and S. Lee, ”A Deep Learning-Based Framework for Offensive Text Detection in Unstructured Data for Heterogeneous Social Media,” *IEEE Access*, vol. 11, pp. 124484-124498, 2023, doi:10.1109/ACCESS.2023.3330081.

[10] H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, and R. Yang, ”Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models,” *Proceedings of the ACM Web Conference 2024 (WWW ’24)*, pp. 2359-2367, May 2024, doi:10.1145/3589334.3645381.

[11] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, F. R. A. Ahmed, A. De, M. A. Khan, and T. M. Ghazal, ”Multimodal Hate Speech Detection in Memes using Contrastive Language-Image Pre-training,” *IEEE Access*, vol. 11, pp. 1-12, 2023, doi:10.1109/ACCESS.2024.3361322.

[12] P. Aggarwal, J. Mehrabanian, W. Huang, O¨ . Alacam, and T. Zesch, ”Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models?,” *arXiv preprint arXiv:2402.04967*, 2024.

[13] T. Do, T. Tran, T. Nguyen, D.-D. Le, T. D. Ngo, ”SignboardText: Text Detection and Recognition in In-the-Wild Signboard Images,” *IEEE Access*, vol. 10, pp. 1-17, 2024, DOI: 10.1109/ACCESS.2024.3395374.

[14] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, ”Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2672–2684, 2019, doi: 10.1109/TPAMI.2019.2934116.

[15] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, ”ABCNet v2: Adaptive Bezier-Curve Network for Real-time End-to-End Text Spotting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 1-17, 2021, DOI: 10.1109/TPAMI.2021.3057374.

[16] W. Wang, E. Xie, X. Li, X. Liu, D. Liang, Z. Yang, T. Lu, and C. Shen, ”PAN++: Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text,” *IEEE Transactions on Pattern Analy- sis and Machine Intelligence*, vol. 43, no. 2, pp. 1-17, 2021, DOI: 10.1109/TPAMI.2021.3057374.

[17] Alan Joseph, Abhinay A K, Dr. Gee Varghese Titus, Anagha Tess B, Adham Saheer, Fabeela Ali Rawther, “Comparative Analysis of Text Classification Models for Offensive Language Detection on Social Media Platforms”, International Journal on Emerging Research Areas (ISSN:2230-9993), vol.04, issue 01, 2024 doi: 10.5281/zenodo.12515626