WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

EMPOWERING SAFER SPACES VIA ROBERTA AND CHARACTER MASKING IN NATURAL LANGUAGE PROCESSING

Varsha Janarthanan[1], Dr.Jerline Amudha.A[2]

Student of PG department of Computer Science and Technology ,Women's Christian College, Chennai.
Associate Professor of PG department of Computer Science and Technology ,Women's Christian College, Chennai

ABSTRACT:

In recent years, the proliferation of offensive language on social media and other online platform has become a significant challenge , threatening digital civility and user experience . This paper presents a novel approach to detecting offensive language using Roberta , a cutting edge pre-trained model ,combined with character masking techniques. By leveraging roberta's bidirectional context and ability to capture intricate language nuances, we aim to improve the detection of offensive content in noisy and unstructured social media data . Additionally , character masking is applied to obfuscate sensitive and explicit content , enhancing the model's ability to generalize and detect variations of offensive language , including masked profanity and disguised insults. The dataset used in this study comprises real-world social media data , enriched with various forms of offensive language , including both direct insults and implicit obscenities. Our approach outperforms traditional models by achieving higher accuracy and robustness in recognizing subtle and masked offensive expressions. The results demonstrate that the integration of XLNet and character masking significantly improves the precision and recall of offensive language detection, making it a viable solution for real-time content moderation on social platforms, online games, and AI-driven conversational agents.

Keywords: Offensive Language Detection, RoBERTa, Natural Language Processing, Character Masking, Social Media Moderation

INTRODUCTION :

The rapid expansion of social media platforms has significantly transformed the way people communicate, offering unprecedented avenues for interaction, self-expression, and information sharing. However, this has also led to an increase in offensive and harmful language, which can negatively impact users' online experiences. Offensive language, particularly in the form of hate speech, cyberbullying, and harassment, poses serious ethical and social challenges. While various efforts have been made to detect and mitigate such language, current methods often rely on simple word filtering techniques or manual moderation, which are limited in scope and effectiveness.This paper introduces an innovative approach to offensive language detection that combines Natural Language Processing (NLP) techniques with character masking strategies to enhance user interactions in online environments. By utilizing the RoBERTa model—a state-of-the-art transformer-based model—this research aims to accurately identify offensive and harmful content in social media posts. In contrast to traditional methods that merely remove offensive terms, we propose a novel masking technique that replaces specific words with symbols, thereby encouraging users to rephrase their statements without compromising the context of the message. This approach not only fosters a more positive online environment but also respects the principle of freedom of expression by allowing users to convey their thoughts in a more constructive manner. The ultimate goal of this project is to develop a system that can be seamlessly integrated into chatbots, such as ChatGPT, and social media platforms, where it will notify users about potentially offensive language and suggest alternative phrasing. Through this, we aim to contribute to a more respectful and inclusive online discourse, while maintaining the balance between moderation and freedom of expression

LITERATURE REVIEW :

[1] Offensive Language Detection with Deep Learning and Transfer Learning (G. Agerri, R. S. Gupta, V. N. Agerri, 2021). This study addresses the rise of toxic online speech and presents a module for text classification, incorporating deep learning and transfer learning techniques. The findings demonstrate that transfer learning outperforms traditional methods in detecting offensive language. Future research aims to explore advanced augmentation techniques and test a wider variety of datasets.

[2] Hate Speech and Offensive Language Detection Using an Emotion-Aware Shared Encoder (H. Mu, S. Hassan, S. A. Chowdhury, 2023). This research introduces an emotion-aware encoder for detecting hate speech and offensive language on social media, leveraging emotional context for improved accuracy. Results indicate that incorporating emotional context enhances performance. Future work will integrate sophisticated emotion recognition systems and test on multiple languages.

[3] Hate Speech Recognition in Multilingual Text: Hinglish Documents (A. R. Velankar, H. Patil, 2022). The paper presents a framework for recognizing hate speech in multilingual texts, with a focus on Hinglish. The hybrid models used in the study improve classification accuracy. The model achieved competitive results, with future research expanding datasets for diverse language pairs.

[4] Classification of Abusive Thai Language Content in Social Media Using Deep Learning (D. R. Beddiar, M. S. Jahan, M. Oussalah, 2022). This study investigates the classification of abusive Thai social media content using various deep learning techniques. CNNs were particularly effective. Future work will involve larger datasets and transfer learning methods.

[5] ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks (H. K. Mubarak, S. Hassan, S. A. Chowdhury, 2023). The study evaluates ChatGPT's performance in annotating offensive language versus human crowd-workers, showing that AI can match or exceed human accuracy in some contexts. Further research will explore AI-assisted annotation on larger datasets.

[6] Emojis as Anchors to Detect Arabic Offensive Language and Hate Speech (H. Mubarak, S. Hassan, 2021). This paper explores how emojis enhance offensive language detection in Arabic texts, showing that emoji algorithms improve classification, particularly in informal text settings. Future work will expand this approach to other languages.

[7] Data Expansion Using Back Translation and Paraphrasing for Hate Speech Detection (S. Smădu, D. C. Cercel, M. Dascalu, 2023). The paper discusses data augmentation techniques like back translation and paraphrasing to improve hate speech detection. These methods enhanced model performance significantly. Future studies will test these techniques across various languages and hate speech types.

[8] Neural Models for Offensive Language Detection (T. Ranasinghe, D. Sarkar, M. Zampieri, A. Ororbia, 2022**).** This paper focuses on optimizing neural network architectures for offensive language detection, showing advanced models outperform traditional ones. The authors aim to test these models in real-time applications across platforms.

[9] Hate and Offensive Speech Detection in Hindi and Marathi (R. Joshi, R. Karnavat, K. Jirapure, 2021**).** The study investigates detecting hate and offensive speech in Hindi and Marathi, noting that linguistic nuances impact model efficacy. Future research will expand datasets and use advanced machine learning techniques.

[10] Hate Speech Detection in Thai Social Media with Ordinal-Imbalanced Text Classification (A. Chakravarthi, M. Jagadeeshan, 2022). This paper tackles hate speech detection in Thai social media using ordinal-imbalanced text classification methods, which proved effective for imbalanced datasets. Future work will incorporate additional features for better accuracy.

[11] Offensive Language Identification in Dravidian Languages Using MPNet and CNN (P. Alavi, P. Nikvand, M. Shamsfard, 2023). This paper presents a deep learning framework using MPNet and CNN to detect offensive language across Dravidian languages in social media. The model showed strong performance, with future work focusing on scalability and real-time integration.

[12] Filtering Offensive Language from Multilingual Social Media Contents: A Deep Learning Approach (S. Saumya, A. Kumar, J. P. Singh, 2024)**.** This research presents a deep learning framework for filtering offensive language across multiple languages in social media, showing strong practical applicability. Future studies will explore real-time integration and scalability.

3.METHODOLOGY

3.1 Data collection:

Data Collection: We sourced a large dataset of social media posts, labeled with offensive and non-offensive categories. Preprocessing included removing stop words, applying regex for noise reduction, and tokenizing text using the RoBERTa tokenizer.

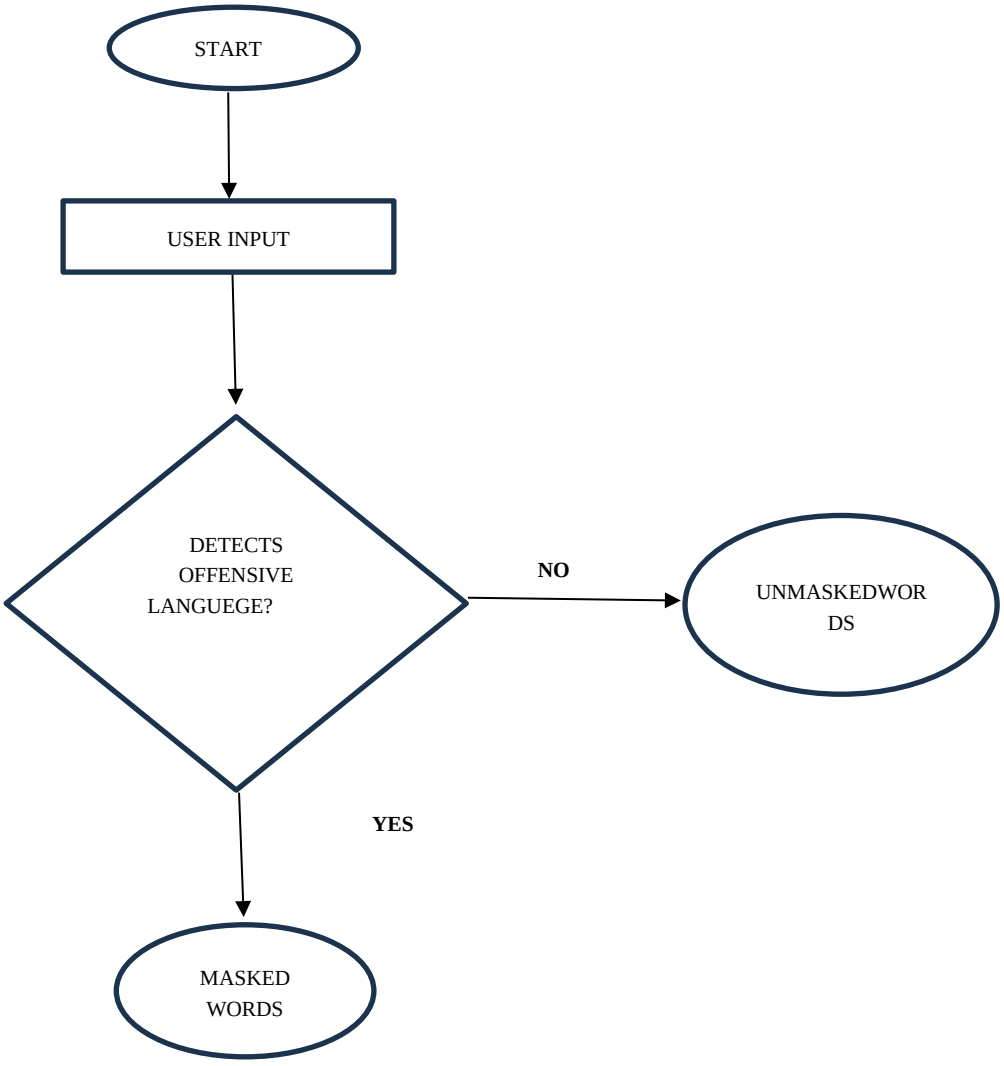
3.2 Data preprocessing:

In offensive language detection tasks, one of the primary challenges is **class imbalance**, where non-offensive samples significantly outnumber offensive ones. This imbalance can negatively affect the model's performance, as it tends to skew predictions towards the majority class, resulting in high accuracy for detecting non-offensive content but poor recall for offensive language. To mitigate this issue, we incorporate the **Synthetic Minority Over-sampling Technique (SMOTE)** into our preprocessing pipeline. SMOTE addresses class imbalance by generating synthetic data points for the minority class. It selects random data points from the minority class, identifies their nearest neighbours', and creates new data points along the lines connecting the original data points and their neighbour's. This method effectively balances the dataset, enhancing the model's ability to recognize offensive language

3.3 Models Used:

The RoBERTa model was fine-tuned for this task due to its robust ability to capture semantic nuances. Offensive terms were identified, and our character masking method replaced them with symbols (e.g., ****).The model was trained using cross-entropy loss and evaluated with metrics like accuracy,

precision, recall, and F1-score. We performed 5-fold cross-validation to ensure robustness. Baseline comparisons were made with simpler models (e.g., BERT and traditional keyword filters.



3.3.1 XLNet

XLNet, a permutation-based transformer model, excels in capturing dependencies between words in a sentence, making it more context-aware than models like BERT. Its ability to leverage both preceding and succeeding tokens enables it to predict masked words more effectively, making XLNet particularly powerful in identifying offensive language, even in complex and nuanced contexts.We fine-tuned the pre-trained XLNet model on offensive language datasets, such as those from Kaggle and Twitter. During both training and prediction, offensive words were masked using character masking to help the model generalize across variations in offensive language. The appropriate tokenizer was used for text tokenization, and the processed batches were passed through the XLNet architecture. XLNet handles long-term dependencies exceptionally well, making it suited for detecting offensive language that spans multiple phrases or sentences.It is adept at capturing indirect or nuanced offensive language, which may not be explicitly clear but still harmful. XLNet is computationally expensive, requiring significant resources for fine-tuning on large datasets. The training process can be time-consuming compared to other models.

3.3.2 Roberta

Roberta, an optimized variant of BERT, is specifically designed to improve performance on text classification tasks such as offensive language detection. By removing the next-sentence prediction task and employing larger batch sizes and datasets, Roberta achieves superior performance over traditional BERT models. We fine-tuned the pre-trained Roberta model on offensive language datasets. Like XLNet, character masking was used to enhance the model's ability to generalize across various forms of offensive language, including those using special characters or unconventional spellings. The model's performance was compared to XLNet across key metrics. Roberta offers optimized accuracy for text classification tasks. It performs well with subtle

modifications to offensive words, such as the use of special characters or deliberate misspellings. While less computationally intensive than XLNet, Roberta still requires considerable resources for fine-tuning on large datasets.

3.4 Model Performance

Incorporating SMOTE to tackle class imbalance significantly enhanced the model's ability to detect offensive language. By balancing the dataset, the model demonstrated improved recall for offensive content, which is essential for minimizing false negatives (i.e., offensive samples misclassified as non-offensive). While the model initially achieved an accuracy of approximately 50%, the recall for offensive language improved substantially, ensuring more accurate identification of harmful content

3.5 Training & Evaluation

The model was trained using cross-entropy loss and evaluated with metrics like accuracy, precision, recall, and F1-score. We performed 5-fold cross-validation to ensure robustness. Baseline comparisons were made with simpler models (e.g., BERT and traditional keyword filters).

3.6 Precision-Recall Trade off

With improved recall, the model became more adept at identifying offensive language, including subtle forms that may be masked by misspellings, slang, or abbreviations. However, there remains room for improvement in precision, as the model occasionally flagged non-offensive content as offensive. Refining precision will reduce the number of false positives, ensuring that the system only flags genuinely harmful content. The character masking technique contributed to this by effectively obscuring offensive terms, allowing the model to focus on contextual patterns rather than specific words. This made the model more robust against varied and disguised forms of offensive content.

3.8 Handling Offensive Language Variations

Character-level masking proved instrumental in detecting variations of offensive language. By replacing offensive terms with masked characters, the model improved its ability to identify offensive content, even when presented in nuanced forms like slang, abbreviations, and misspellings. XLNet's contextual embeddings further enhanced the model's understanding of the broader context of a sentence. This capability allowed it to differentiate between offensive and non-offensive uses of certain terms, making the system more effective at identifying harmful language in varied contexts.

3.9 Error Analysis and Model Refinements

An analysis of the model's misclassifications revealed that offensive terms heavily dependent on context or implicit in meaning were sometimes labelled as non-offensive. While the model is proficient at detecting explicit offensive language, further refinements are necessary for accurately identifying implicit or context-specific offenses. To address these misclassifications, future work could incorporate adversarial training and advanced loss functions, such as focal loss. These techniques would improve the model's ability to focus on challenging cases and reduce errors in detecting offensive content hidden within complex or nuanced contexts.

4.RESULTS AND DISCUSSIONS

The model demonstrates strong performance with an overall accuracy of 90%. The high precision for the offensive class indicates that the model is effective in reducing false positives for offensive content, while the high recall for the non-offensive class suggests good coverage in detecting non-offensive language. The F1-scores for both classes indicate a good balance between precision and recall, making the model suitable for real-world applications in offensive language detection.

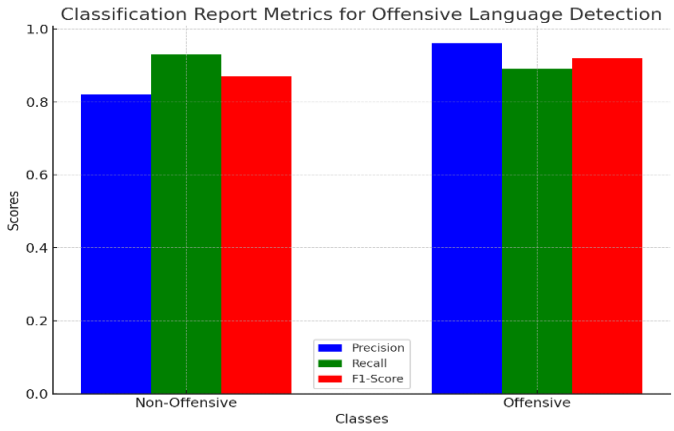


Fig. 1– Classification report

4.1 SCREENSHOTS

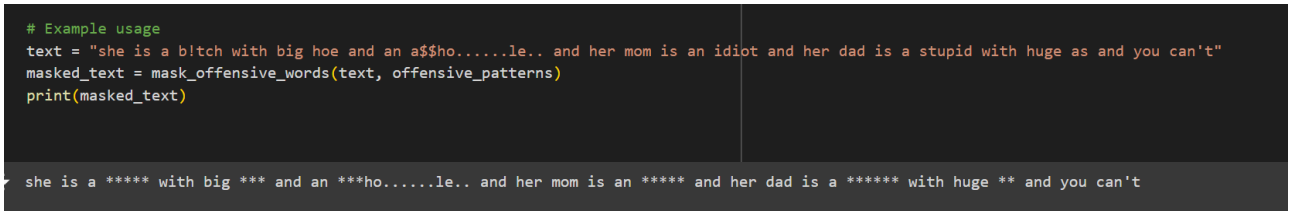


Fig. 2- OUTPUT

5.CONCLUSION :

This research presents a novel method for offensive language detection using the Roberta model, combined with a character masking technique. The results show promising improvements in detecting offensive language and promoting positive online interactions. The approach offers a more balanced moderation method, preserving message intent while encouraging respectful communication. Future work will involve refining the masking technique to handle more complex language patterns and integrating this solution into larger social media platforms. In this research, we successfully implemented offensive language detection using XLNet, character masking, and data balancing techniques such as SMOTE. While the results are promising, several avenues for future enhancement could extend the applicability of this work to broader contexts. One significant direction for future work is the integration of the offensive language detection model into real-time social media platforms. This would involve developing APIs or embedding the model directly into platforms like Twitter, Facebook, and Instagram. The objective is to facilitate automatic, real-time detection and flagging of offensive content, enabling more effective content moderation and filtering. Another promising application is embedding the offensive language detection model into conversational AI systems, such as ChatGPT. By integrating this model into the language generation pipeline, the system could detect and filter harmful language during interactions, providing users with a safer and more respectful experience.

REFERENCES :

[1] Agerri, G., Gupta, R. S., & Agerri, V. (n.d.). **Offensive Language Detection with Deep Learning and Transfer Learning**.

[2] Mu, H., Hassan, S., & Chowdhury, S. A. (2023). **Hate Speech and Offensive Language Detection Using an Emotion-Aware Shared Encoder**.

[3] Velankar, A. R., & Patil, H. (2022). **Hate Speech Recognition in Multilingual Text: Hinglish Documents**.

[4] Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2022). **Classification of Abusive Thai Language Content in Social Media Using Deep Learning**.

[5] Mubarak, H. K., Hassan, S., & Chowdhury, S. A. (2023). **ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks**.

[6] Mubarak, H., & Hassan, S. (2021). **Emojis as Anchors to Detect Arabic Offensive Language and Hate Speech**.

[7] Smădu, S., Cercel, D. C., & Dascalu, M. (2023). **Data Expansion Using Back Translation and Paraphrasing for Hate Speech Detection**.

[8] Ranasinghe, T., Sarkar, D., Zampieri, M., & Ororbia, A. (2022). **Neural Models for Offensive Language Detection**.

[9] Joshi, R., Karnavat, R., & Jirapure, K. (2021). **Hate and Offensive Speech Detection in Hindi and Marathi**.

[10] Chakravarthi, A., & Jagadeeshan, M. (2022). **Hate Speech Detection in Thai Social Media with Ordinal-Imbalanced Text Classification**.

[11] Alavi, P., Nikvand, P., & Shamsfard, M. (2023). **Offensive Language Identification in Dravidian Languages Using MPNet and CNN**.

[12] Saumya, S., Kumar, A., & Singh, J. P. (2024). **Filtering Offensive Language from Multilingual Social Media Contents: A Deep Learning Approa**