International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# Systematic Review on Offensive Language Detection of Multilingual Texts using NLP and Machine Learning Techniques

**Roopa G K**

Assistant Professor, Department of Computer Science &Engg, VCET, Puttur
Email: *roopa.rkumbady[at]gmail.com*

**Abstract:** ***The widespread usage of social media and online platforms has given rise to a substantial increase in offensive language. Detecting offensive content is crucial to maintaining a healthy online environment and protecting users from harm. Natural Language Processing (NLP) and Machine Learning (ML) techniques have shown promise in addressing this challenge, especially in the context of multilingual texts. This paper presents a systematic review of the existing literature on offensive language detection in multilingual texts, focusing on the NLP and ML methodologies utilized, dataset characteristics, evaluation metrics, and performance comparisons. The review aims to provide an overview of the state-of-the-art techniques, identify key challenges, and suggest future research directions in this area.***

**Keywords:** Natural Language Processing (NLP), Machine Learning (ML), multilingual texts

## 1. Introduction

The proliferation of social media platforms and online communication channels has provided an unprecedented avenue for individuals to express their thoughts and opinions. However, this digital revolution has also brought to light the dark side of online discourse - the rampant use of offensive language, hate speech, and harmful content. Offensive language not only poses a threat to the emotional well-being of users but also undermines the quality of online interactions and fosters a toxic environment.

Efficiently detecting offensive language in multilingual texts is a challenging task, as it requires understanding the intricacies of various languages, cultural contexts, and linguistic nuances. Traditional rule-based methods and keyword filtering are inadequate in dealing with the dynamic and context-dependent nature of offensive content. To address this issue, researchers have turned to Natural Language Processing (NLP) and Machine Learning (ML) techniques, which have shown considerable promise in automating the identification of offensive language across different languages and settings.

The motivation behind this systematic review is to comprehensively analyze the state-of-the-art offensive language detection approaches that employ NLP and ML techniques for multilingual texts. By examining existing literature and research, we aim to provide a comprehensive overview of the methodologies used, assess the performance of various models, and identify the key challenges that researchers encounter in this domain. Furthermore, this review seeks to suggest potential research directions to improve the accuracy and generalizability of offensive language detection systems.

### Objectives

The main objectives of this systematic review are as follows:

1) Provide an extensive overview of the NLP and ML techniques utilized for offensive language detection in multilingual texts.
2) Investigate and assess the characteristics of available multilingual offensive language datasets, including their diversity, size, and language distribution.
3) Analyze the performance of state-of-the-art models and techniques for offensive language detection across different languages and compare their effectiveness.
4) Identify and discuss the challenges and limitations faced by existing approaches, such as handling data imbalance and cross-linguistic ambiguity.
5) Propose future research directions to enhance the capabilities and robustness of offensive language detection systems, including the incorporation of context, pragmatics, and ethical considerations.

## 2. Offensive Language Detection: Overview

Offensive language detection is a critical area of research with significant implications for online communities, platforms, and users. The combination of NLP and machine learning techniques has led to substantial progress in detecting offensive content. However, challenges such as multilingual variations, context understanding, and evolving language usage require continuous research and innovation to develop robust and culturally sensitive offensive language detection systems.

### 2.1 Definition and Types of Offensive Language

Offensive language encompasses a wide range of harmful content, including hate speech, profanity, cyber bullying, harassment, and discriminatory remarks. These expressions can target individuals or groups based on attributes such as

race, religion, ethnicity, gender, or other personal characteristics. Offensive language can be explicit or implicit, making its detection challenging, especially in multilingual contexts.

### 2.2 Importance of Offensive Language Detection

The impact of offensive language in online spaces is far-reaching and can lead to severe consequences for individuals and communities. It can cause emotional distress, promote hatred and violence, incite discrimination, and stifle open dialogue. Offensive content also poses risks for online platforms, leading to reputational damage, legal liabilities, and potential user churn. Therefore, effective offensive language detection is crucial for content moderation, enforcing community guidelines, and safeguarding the well-being of users.

### 2.3 Approaches to Offensive Language Detection:

**2.3.1 Traditional Rule-Based Methods:** Early efforts in offensive language detection involved manual creation of rule-based systems that relied on predefined lists of offensive words and patterns. While these approaches could handle straightforward cases, they lacked the ability to capture the context and evolving nature of offensive language.

**2.3.2 Machine Learning Techniques:** With the advancements in machine learning, data-driven approaches gained popularity for offensive language detection. Supervised learning methods, such as Support Vector Machines (SVM), Naive Bayes, and deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have demonstrated promising results.

**2.3.3 Natural Language Processing (NLP) Techniques:** NLP techniques play a vital role in understanding and representing language in offensive language detection. Tokenization, text preprocessing, and language representation using word embeddings (e.g., Word2Vec, GloVe, and fastText) help in transforming raw text into suitable inputs for machine learning models.

**2.3.4 Multilingual Challenges:** Detecting offensive language in multilingual texts introduces additional complexities due to variations in language structure, culture, and expressions. NLP models must account for different character sets, writing systems, and linguistic features across languages.

**2.3.5 Context and Ambiguity:** Offensive language detection often requires considering the context of words and phrases to determine their intended meaning. Certain terms may be offensive in some contexts but neutral or even positive in others. This ambiguity poses a significant challenge in achieving accurate detection.

## 3. NLP Techniques for Offensive Language Detection

NLP (Natural Language Processing) techniques play a vital role in offensive language detection by processing and understanding textual data. These techniques transform raw text into suitable representations, enabling machine learning models to learn patterns and make accurate predictions.

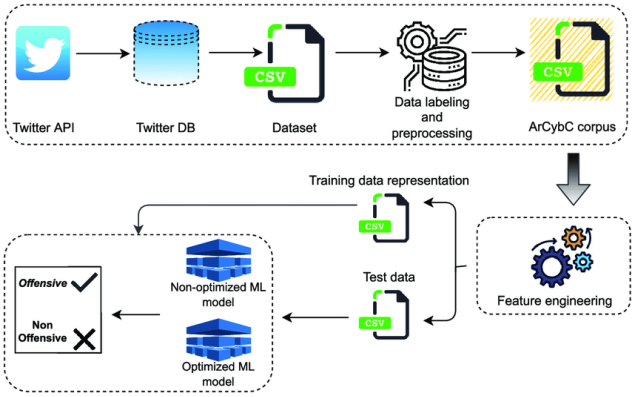


**Figure1:** Block Diagram of NLP Techniques

### 3.1 Tokenization

Tokenization is the process of breaking a text into smaller units called tokens. Tokens can be words, subwords, or characters, depending on the granularity required. Tokenization helps to structure the text, making it easier for subsequent NLP tasks.

### 3.2 Text Preprocessing

Text preprocessing involves cleaning and normalizing the text data to remove noise and irrelevant information. Common preprocessing steps include converting text to lowercase, removing punctuation, expanding contractions, and handling special characters.

### 3.3 Word Embeddings

Word embeddings are dense vector representations of words in a continuous vector space. Techniques like Word2Vec, GloVe, and fastText create word embeddings that capture semantic relationships between words. These embeddings help in understanding the meaning and context of words in offensive language detection tasks.

### 3.4 Language Representations

For offensive language detection in multilingual texts, language representations are essential to handle diverse languages and linguistic characteristics. Multilingual word embeddings and contextual language models like BERT (Bidirectional Encoder Representations from Transformers) are used to encode text and capture cross-lingual semantic information.

### 3.5 Feature Extraction:

Feature extraction involves extracting informative features from the text that can be used as input to machine learning

models. N-grams, Bag-of-Words (BoW), and TF-IDF (Term Frequency-Inverse Document Frequency) are traditional feature extraction methods. In contrast, neural networks can automatically learn relevant features from the text data.

### 3.6 Deep Learning Architectures

Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown significant success in offensive language detection. CNNs are effective in capturing local patterns, while RNNs can capture sequential dependencies in text.

### 3.7 Transfer Learning and Multilingual Models

Transfer learning techniques leverage pre-trained language models on large-scale datasets to improve offensive language detection tasks with limited data. Models like BERT, GPT (Generative Pre-trained Transformer), and XLM-R (Cross-lingual Language Model) have achieved state-of-the-art results in multilingual offensive language detection.

### 3.8 Contextual Information

Understanding the context in which offensive words or phrases appear is crucial for accurate detection. Contextual analysis, such as examining surrounding words or sentences, helps in disambiguating potentially offensive content from neutral expressions.

### 3.9 Ensembling

Ensemble methods combine multiple models to improve predictive performance. Combining the outputs of different offensive language detection models, either by voting or averaging, can enhance overall accuracy and reduce false positives.

By leveraging these NLP techniques, offensive language detection models can effectively analyze text data and distinguish between offensive and non-offensive content, promoting a safer and more respectful online environment. Continuous research and advancements in NLP are essential to address the evolving nature of offensive language and adapt to the linguistic diversity of online communication platforms.

## 4. Machine Learning Techniques for Offensive Language Detection:

### 4.1 Supervised Learning Approaches

Supervised learning is a popular approach in offensive language detection, where the model is trained on labeled data, meaning the input text is paired with corresponding offensive or non-offensive labels. The model learns from these examples to make predictions on unseen data. Common supervised learning algorithms used in offensive language detection include:

- **Support Vector Machines (SVM):** SVM is a binary classification algorithm that aims to find the optimal hyperplane that separates offensive and non-offensive instances in the feature space.
- **Naive Bayes:** Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that features are conditionally independent given the class label, making it computationally efficient for text classification tasks.
- **Logistic Regression:** Logistic Regression is another binary classifier that models the probability of a text belonging to an offensive class.
- **Deep Learning Models:** Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have also been used in supervised learning for offensive language detection. CNNs are effective in capturing local patterns, while RNNs can model sequential dependencies in text.

### 4.2 Unsupervised and Semi-Supervised Learning Approaches

Unsupervised learning techniques do not rely on labeled data but instead aim to discover patterns and structures within the data. Semi-supervised learning methods, on the other hand, leverage a small amount of labeled data along with a larger pool of unlabeled data to improve model performance. In the context of offensive language detection, these approaches are less commonly used, but they offer potential advantages in certain scenarios:

- **Clustering:** Unsupervised clustering algorithms group similar texts together based on their features, which can help identify potential clusters of offensive content.
- **Topic Modeling:** Topic modeling techniques like Latent Dirichlet Allocation (LDA) can uncover latent topics in a corpus, enabling the identification of topics that might be related to offensive language.
- **Self-training**: Self-training is a semi-supervised approach that involves training a model with a small labeled dataset and then using that model to predict labels for unlabeled data. The newly labeled data is combined with the original labeled data, and the model is retrained iteratively.

### 4.3 Ensemble Techniques

Ensemble techniques combine multiple base models to improve overall performance. In the context of offensive language detection, ensemble methods are particularly useful for reducing false positives and improving robustness. Some commonly used ensemble techniques include:

- **Voting:** In a voting ensemble, multiple offensive language detection models (e.g., SVM, Naive Bayes, and CNN) make predictions on the same input text, and the most frequent prediction becomes the final output.
- **Stacking:** Stacking involves training a meta-model that takes the outputs of individual offensive language detectors as its input features. The meta-model then makes the final prediction based on these combined outputs.
- **Bagging:** Bagging (Bootstrap Aggregating) involves training multiple instances of the same offensive

language detection model on different subsets of the training data and combining their predictions.

- **Boosting:** Boosting is an ensemble technique where models are trained sequentially, with each new model focusing on correcting the errors of the previous models.

Ensemble techniques help mitigate the weaknesses of individual models and can lead to improved offensive language detection performance, making them valuable tools for handling the intricacies of offensive content classification.

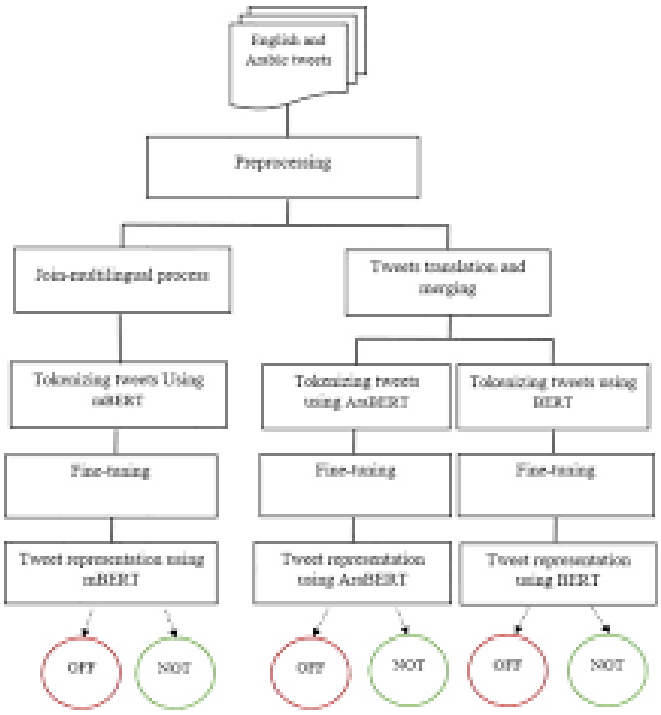


**Figure 2:** Machine Learning Techniques for Offensive Language Detection

## 5. Multilingual Text Datasets for Offensive Language Detection

### 5.1 Compilation of Datasets

The availability of high-quality and diverse multilingual datasets is essential for training and evaluating offensive language detection models across different languages. Researchers have made efforts to compile datasets from various sources to facilitate research in this area. Some common multilingual offensive language datasets include:

- **Wikipedia Detox**: Wikipedia Detox datasets contain comments from different language editions of Wikipedia, labeled for toxicity, aggression, and personal attacks.
- **OpenSubtitles:** The OpenSubtitles dataset comprises subtitle data from movies and TV shows, available in multiple languages, and annotated for offensive content.
- **HASOC** (Hate Speech and Offensive Content): The HASOC dataset covers offensive content in Hindi, German, and English social media posts collected from Twitter and Facebook.
- **OffensEval (SemEval):** The OffensEval dataset is part of the SemEval shared task on offensive language detection and includes multilingual data from Twitter and other social media platforms.
- **TRAC** (Trolling, Aggression, and Cyberbullying): The TRAC dataset consists of multilingual social media comments annotated for aggression, cyberbullying, and trolling.
- **PolEval**(Polish Offensive Language Dataset): The PolEval dataset focuses on offensive language in Polish social media posts.

### 5.2 Characteristics of Multilingual Offensive Language Datasets:

The characteristics of multilingual offensive language datasets can significantly influence the performance and generalizability of offensive language detection models. Some key characteristics to consider include:

- **Language Diversity:** Multilingual datasets should encompass a wide range of languages to cater to the linguistic diversity of online communication platforms.
- **Annotation Consistency:** Ensuring consistent and reliable annotation of offensive language across languages is crucial for fair evaluation and comparison of models.
- **Class Imbalance**: Offensive language datasets often suffer from class imbalance, where the number of offensive instances is significantly smaller than non-offensive instances. Addressing class imbalance is essential to prevent bias in model performance.
- **Text Length and Complexity**: Datasets should contain texts of varying lengths and complexities to capture the challenges posed by short messages, long paragraphs, and different writing styles.
- **Real-World Context:** Offensive language detection models need to be trained on real-world data that closely represents the offensive content found on social media and other online platforms.

### 5.3 Evaluation and Comparison of Datasets

To evaluate and compare offensive language detection models effectively, it is essential to standardize evaluation metrics and experimental setups. Common evaluation metrics include precision, recall, F1-score, accuracy, and area under the receiver operating characteristic curve (AUC-ROC).

Researchers often employ cross-validation techniques to mitigate overfitting and ensure robust evaluation. Additionally, it is crucial to establish baseline performance using well-established models and techniques on each dataset. Comparative analysis of various datasets can shed light on the challenges posed by different languages and the effectiveness of models across diverse linguistic contexts.

## 6. Evaluation Metrics and Benchmarking

### 6.1 Common Evaluation Metrics

Evaluation metrics are essential for quantifying the performance of offensive language detection models. Various metrics are used to assess the model's effectiveness

in correctly classifying offensive and non-offensive texts. Some common evaluation metrics include:

**Precision:** Precision measures the proportion of true positive instances (correctly classified offensive texts) out of all instances predicted as offensive. High precision indicates a low false positive rate.

**Recall (Sensitivity or True Positive Rate):** Recall measures the proportion of true positive instances out of all actual offensive instances. High recall indicates a low false negative rate.

**F1-score:** The F1-score is the harmonic mean of precision and recall and provides a balanced measure of model performance.

**Accuracy:** Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) out of the total instances. However, accuracy may be misleading in imbalanced datasets, where the number of offensive and non-offensive instances differs significantly.

**Area Under the Receiver Operating Characteristic curve (AUC-ROC):** AUC-ROC measures the model's ability to distinguish between offensive and non-offensive instances across different classification thresholds. It is particularly useful in imbalanced datasets.

**Area Under the Precision-Recall curve (AUC-PR):** AUC-PR quantifies the trade-off between precision and recall, especially in imbalanced datasets.

### 6.2 Benchmarking of State-of-the-Art Models:

Benchmarking involves comparing the performance of offensive language detection models against well-established baseline models and state-of-the-art approaches. Benchmarking allows researchers to assess the progress made in the field and identify areas for improvement. Common benchmarking practices include:

- **Utilizing Standard Datasets:** Researchers use publicly available datasets with well-defined labels to ensure fair comparison between different models.
- **Cross-Validation:** Cross-validation techniques, such as k-fold cross-validation, are employed to avoid overfitting and provide a more robust evaluation.
- **Reproducibility:** To ensure the reproducibility of results, researchers often share their code, model architecture, and hyperparameters.
- **Reporting Results:** Results should be reported with clear details on the evaluation metrics, dataset characteristics, and any modifications made to the model.

### 6.3 Challenges in Evaluation:

Evaluating offensive language detection models presents several challenges that need to be addressed to obtain reliable and meaningful results:

**Data Imbalance:** Offensive language datasets are often imbalanced, with a significant class skew between offensive and non-offensive instances. This imbalance can lead to biased evaluation results, and specific evaluation metrics (e.g., accuracy) may not reflect the model's true performance.

**Multilingual Evaluation:** Evaluating models on multilingual datasets introduces additional complexities due to language variations and linguistic differences**.** Proper handling of language-specific nuances is crucial for fair evaluation.

**Contextual Understanding:** Offensive language detection requires understanding the context in which offensive words are used. Ensuring the evaluation data includes diverse contexts is essential for assessing model performance accurately.

**Bias and Fairness:** Offensive language detection models can be sensitive to biased training data, leading to biased predictions. Evaluating models for fairness and bias is vital to avoid discriminatory outcomes.

**Cross-Domain Generalization:** Models trained on specific datasets may not generalize well to new domains, making cross-domain evaluation essential for assessing real-world applicability.

Addressing these challenges in evaluation is vital to develop robust offensive language detection models that perform effectively across different languages, contexts, and domains. Researchers must continuously strive for improved evaluation practices and benchmarking standards to advance the field and promote responsible and ethical use of offensive language detection technology.

## 7. Performance Comparison of NLP and ML Techniques

### 7.1 Comparative Analysis of NLP Techniques:

NLP techniques have played a crucial role in advancing offensive language detection systems. Comparative analysis of NLP techniques used in offensive language detection involves assessing the strengths and weaknesses of various methodologies. Some aspects of the analysis include:

**Tokenization and Text Preprocessing**: The effectiveness of different tokenization and text preprocessing approaches in handling offensive content, especially in languages with complex linguistic structures, is evaluated.

**Word Embeddings and Language Representations:** Comparative analysis explores the impact of using different word embeddings and language representations in capturing semantic relationships and context for offensive language detection.

**Feature Extraction and Selection:** The performance of various feature extraction methods (e.g., N-grams, BoW, and TF-IDF) in representing offensive content is compared to understand their effectiveness in different languages.

**Deep Learning Architectures:** The performance of CNNs, RNNs, and transformer-based models (e.g., BERT, GPT) in

offensive language detection is assessed to identify the most suitable architecture for different language contexts.

### 7.2 Comparative Analysis of ML Techniques:

Comparative analysis of ML techniques for offensive language detection involves evaluating the performance of various algorithms and models. Key aspects of the analysis include:

**Supervised Learning Approaches:** The performance of different supervised learning algorithms, such as SVM, Naive Bayes, and logistic regression, is compared to identify the most effective classifiers for offensive language detection.

**Unsupervised and Semi-Supervised Learning Approaches:** The performance of unsupervised and semi-supervised techniques, such as clustering and self-training, is evaluated to understand their usefulness in scenarios with limited labeled data.

**Ensemble Techniques:** The effectiveness of ensemble methods, such as voting, stacking, bagging, and boosting, is analyzed to determine the benefits of combining multiple models.

### 7.3 Cross-Linguistic Performance Variations:

Offensive language detection models often encounter variations in performance across different languages due to linguistic differences, cultural context, and data availability. Comparative analysis of cross-linguistic performance involves:

**Language-specific Challenges:** Identifying language-specific challenges that affect model performance, such as linguistic ambiguity, morphological differences, and the availability of labeled data.

**Transfer Learning:** Investigating the effectiveness of transfer learning techniques, where models pre-trained on one language are fine-tuned for another language to handle cross-linguistic variations.

**Multilingual Models:** Assessing the performance of multilingual models, such as multilingual word embeddings and transformer-based models, to understand their ability to handle offe**nsive language detection across diverse languages.**

Comparative analysis of NLP and ML techniques helps researchers and practitioners identify the most effective methodologies for offensive language detection in different language contexts. Understanding the strengths and weaknesses of various approaches allows for the development of more robust and accurate models that can address the challenges posed by offensive content across various languages and cultures. Additionally, cross-linguistic performance variations provide valuable insights into the generalization capabilities of offensive language detection models, enabling improvements in their real-world applications.

## 8. Challenges and Limitations

### 8.1 Data Imbalance

Data imbalance is a common challenge in offensive language detection, where the number of offensive instances is significantly smaller than non-offensive instances. This imbalance can lead to biased model training and skewed evaluation results. Addressing data imbalance is essential to ensure that offensive language detection models do not favor the majority class and can accurately identify offensive content. Some strategies to mitigate data imbalance include:

**Resampling Techniques:** Using oversampling (adding more instances of the minority class) or undersampling (removing instances from the majority class) to balance the dataset.

**Class Weighting:** Assigning higher weights to the minority class during model training to give it more importance.

**Data Augmentation:** Generating synthetic offensive instances to increase the representation of the minority class.

**Cost-Sensitive Learning:** Modifying the learning algorithm to penalize misclassifications of the minority class more than the majority class.

### 8.2 Handling Multilingual Ambiguity:

Offensive language detection in multilingual contexts introduces challenges related to language-specific ambiguity and variations. The same words or phrases may have different meanings or offensive connotations in different languages or cultural contexts. Handling multilingual ambiguity requires specialized techniques to:

**Develop Language-Specific Models:** Creating separate models for each language can help capture language-specific nuances and improve model performance.

**Leveraging Multilingual Models:** Using pre-trained multilingual language models like multilingual BERT or XLM-R can help the model generalize across different languages and improve performance in multilingual settings.

**Cross-Lingual Transfer Learning:** Transfer learning techniques allow models trained on one language to be adapted to another, enabling better handling of cross-linguistic ambiguity.

### 8.3 Out-of-Domain Generalization

Out-of-domain generalization refers to the ability of offensive language detection models to perform well on data from domains different from their training data. Models trained on specific datasets may struggle to generalize to new and diverse contexts. To improve out-of-domain generalization, researchers must:

**Use Diverse Training Data:** Training offensive language detection models on diverse datasets that cover a wide range

of offensive content is essential to improve their ability to handle various contexts.

**Incorporate Domain Adaptation Techniques:** Techniques like domain adaptation and transfer learning help models adapt to new domains with limited labeled data.

**Fine-Tuning on Target Domains:** Fine-tuning models on target domain data helps align the model's performance with the specific characteristics of the new domain.
**Ensemble Methods:** Combining models trained on different domains using ensemble techniques can enhance overall model performance and generalization.

Addressing these challenges and limitations is crucial to building effective and robust offensive language detection systems that can accurately detect offensive content across different languages, contexts, and domains. Ongoing research and advancements in NLP and machine learning will play a significant role in improving the capabilities of these models and fostering a safer and more respectful online environment.

## 9. Future Research Directions:

### 9.1 Multimodal Offensive Language Detection

Current offensive language detection models primarily focus on text data, but offensive content can also be conveyed through images, videos, and audio. Future research should explore the integration of multimodal approaches, combining text, images, and audio data to enhance offensive language detection accuracy. Multimodal models can capture additional contextual cues and non-textual features, leading to a more comprehensive understanding of offensive content across different media types.

### 9.2 Incorporating Context and Pragmatics

The meaning of offensive language heavily depends on the surrounding context and pragmatic factors. Future research should concentrate on developing models that can effectively understand and incorporate context to improve offensive language detection accuracy. Techniques like discourse analysis, conversational context, and speaker intent recognition can aid in capturing the subtleties of offensive language and discerning between offensive and non-offensive use cases.

### 9.3 Ethical and Sociocultural Considerations

As offensive language detection technology becomes more prevalent, researchers must address the ethical implications of its use. Ensuring fairness, transparency, and mitigating bias are critical considerations. Future research should focus on developing models that are sensitive to cultural nuances and diverse perspectives to avoid inadvertently perpetuating biases or harming certain social groups. Additionally, ethical guidelines for the deployment of offensive language detection systems in real-world applications need to be established to safeguard user rights and privacy.

### 9.4 Development of Benchmark Datasets

The availability of high-quality benchmark datasets is essential for advancing offensive language detection research. Future research should focus on creating more diverse, balanced, and multilingual datasets that reflect the complexities of offensive content in various contexts. The development of standardized evaluation protocols and benchmark datasets will facilitate fair comparison and benchmarking of offensive language detection models, encouraging collaboration and fostering advancements in the field.

### 9.5 Lifelong and Continual Learning

Offensive language evolves rapidly, and models need to adapt to changing language patterns and new offensive terms. Future research should explore lifelong and continual learning techniques to enable offensive language detection models to update and fine-tune themselves with new data over time. This approach ensures that models remain effective and up-to-date in detecting emerging offensive content.

### 9.6 Online and Real-Time Deployment

Efficient and real-time deployment of offensive language detection models is crucial for content moderation on social media platforms and online communication channels. Future research should focus on developing lightweight and scalable models that can be integrated seamlessly into online platforms, enabling real-time detection and response to offensive content.

In conclusion, future research in offensive language detection should explore the integration of multimodal approaches, incorporate context and pragmatics, consider ethical and sociocultural considerations, develop benchmark datasets, enable lifelong learning, and focus on real-time deployment. By addressing these research directions, offensive language detection systems can become more robust, inclusive, and effective inpromoting safer and more respectful online interactions.

## 10. Conclusion

Offensive language detection in multilingual texts using NLP and machine learning techniques is a critical area of research with significant implications for creating safer and more inclusive online spaces. This systematic review explored the various aspects related to offensive language detection, including NLP techniques, machine learning approaches, multilingual datasets, evaluation metrics, and challenges faced by existing models.

NLP techniques, such as tokenization, word embeddings, and language representations, play a crucial role in understanding and processing textual data for offensive language detection. Machine learning approaches, including supervised, unsupervised, and ensemble techniques, have demonstrated promising results in automating the identification of offensive content.

Multilingual offensive language datasets offer valuable resources for training and evaluating models in diverse linguistic contexts. However, challenges such as data imbalance, handling multilingual ambiguity, and out-of-domain generalization must be addressed to ensure accurate and unbiased offensive language detection.

Evaluation metrics and benchmarking are essential for objectively assessing model performance and comparing state-of-the-art approaches. Cross-linguistic performance variations highlight the need for robust and adaptable offensive language detection systems capable of handling diverse languages and cultural contexts.

In conclusion, offensive language detection in multilingual texts using NLP and machine learning techniques is an evolving field with immense potential to promote respectful and inclusive online communication. By continually advancing research and addressing challenges, researchers and practitioners can build sophisticated and culturally sensitive offensive language detection systems, contributing to a safer and more harmonious digital environment for users worldwide.

## References

[1] Kumar, R.; Ojha, A.K.; Malmasi, S.; Zampieri, M. Evaluating Aggression Identification in Social Media. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 16 May 2020; pp. 1–5. [Google Scholar]

[2] Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 75–86. [Google Scholar] [CrossRef][Green Version]

[3] Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F.M.; Rosso, P.; Sanguinetti, M. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 54–63. [Google Scholar] [CrossRef][Green Version]

[4] Sai, S.; Sharma, Y. Towards Offensive Language Identification for Dravidian Languages. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Kyiv, Ukraine, 19–20 April 2021; pp. 18–27. [Google Scholar]

[5] Hettiarachchi, H.; Adedoyin-Olowe, M.; Bhogal, J.; Gaber, M.M. Embed2Detect: Temporally clustered embedded words for event detection in social media. Mach. Learn. 2021. [Google Scholar] [CrossRef]

[6] Akbar, S.Z.; Panda, A.; Kukreti, D.; Meena, A.; Pal, J. Misinformation as a Window into Prejudice: COVID-19 and the Information Environment in India. Proc. ACM Hum.-Comput. Interact. 2021, 4. [Google Scholar] [CrossRef]

[7] Sharma, I. Contextualising Hate Speech: A Study of India And Malaysia. Millenn. J. Int. Stud. 2019, 15, 133–144. [Google Scholar] [CrossRef]

[8] Ranasinghe, T.; Zampieri, M. Multilingual Offensive Language Identification for Low-resource Languages. arXiv 2021, arXiv:2105.05996. [Google Scholar]

[9] Ranasinghe, T.; Zampieri, M. MUDES: Multilingual Detection of Offensive Spans. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, Online, 8–9 June 2021; pp. 144–152. [Google Scholar] [CrossRef]

[10] Solorio, T.; Blair, E.; Maharjan, S.; Bethard, S.; Diab, M.; Ghoneim, M.; Hawwari, A.; AlGhamdi, F.; Hirschberg, J.; Chang, A.; et al. Overview for the First Shared Task on Language Identification in Code-Switched Data. In Proceedings of the First Workshop on Computational Approaches to Code Switching, Doha, Qatar, 25 October 2014; pp. 62–72. [Google Scholar] [CrossRef][Green Version]

[11] Mubarak, H.; Darwish, K.; Magdy, W. Abusive Language Detection on Arabic Social Media. In Proceedings of the First Workshop on Abusive Language, Vancouver, BC, Canada, 4 August 2017; pp. 52–56. [Google Scholar] [CrossRef][Green Version]

[12] Vidgen, B.; Derczynski, L. Directions in abusive language training data, a systematic review: Garbage in, garbage out. PLoS ONE 2021, 15, e0243300. [Google Scholar] [CrossRef]

[13] Kumar, R.; Ojha, A.K.; Malmasi, S.; Zampieri, M. Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM, USA, 25 August 2018; pp. 1–11. [Google Scholar]

[14] Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A.V.; Trancoso, I. Automatic cyberbullying detection: A systematic review. Comput. Hum. Behav. 2019, 93, 333–345. [Google Scholar] [CrossRef]

[15] Malmasi, S.; Zampieri, M. Detecting Hate Speech in Social Media. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, 2–8 September 2017; pp. 467–472. [Google Scholar] [CrossRef]

[16] Malmasi, S.; Zampieri, M. Challenges in Discriminating Profanity from Hate Speech. J. Exp. Theor. Artif. Intell. 2018, 30, 1–16. [Google Scholar] [CrossRef][Green Version]

[17] Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; Pierrehumbert, J. HateCheck: Functional Tests for Hate Speech Detection Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 17 July 2021; pp. 41–58. [Google Scholar] [CrossRef]

[18] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [Google Scholar] [CrossRef]
[19] Ranasinghe, T.; Zampieri, M.; Hettiarachchi, H. BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In Proceedings of the 11th Forum for Information Retrieval, Kolkata, India, 12–15 December 2019. [Google Scholar]
[20] Mandl, T.; Modha, S.; Kumar M, A.; Chakravarthi, B.R. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In Proceedings of the FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, 16–20 December 2020; pp. 29–32. [Google Scholar] [CrossRef]
[21] Mubarak, H.; Rashed, A.; Darwish, K.; Samih, Y.; Abdelali, A. Arabic Offensive Language on Twitter: Analysis and Experiments. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (Virtual), Kyiv, Ukraine, 19 April 2021; pp. 126–135. [Google Scholar]
[22] Pitenis, Z.; Zampieri, M.; Ranasinghe, T. Offensive Language Identification in Greek. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 5113–5119. [Google Scholar]
[23] Çöltekin, Ç. A Corpus of Turkish Offensive Language on Social Media. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6174–6184. [Google Scholar]
[24] Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451. [Google Scholar] [CrossRef]
[25] Risch, J.; Krestel, R. Bagging BERT Models for Robust Aggression Identification. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 16 May 2020; pp. 55–61. [Google Scholar]
[26] Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; Patel, A. Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In Proceedings of the 11th Forum for Information Retrieval Evaluation, Kolkata, India, 12–15 December 2019; pp. 14–17. [Google Scholar] [CrossRef]
[27] Chakravarthi, B.R.; Priyadharshini, R.; Jose, N.; Mandl, T.; Kumaresan, P.K.; Ponnusamy, R.; Hariharan, R.L.; McCrae, J.P.; Sherly, E. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Kyiv, Ukraine, 19 April 2021; pp. 133–145. [Google Scholar]

## Author Profile

**Roopa G K** received the B.E degree in CSE from KVG College of Engineering, Sullia and M.Tech. degree in CSE from NMAMIT, Nitte. She is currently pursuing her PhD from NITK, Surathkal under the guidance of Prof. Santhi Thilagam in the area of NLP. Her areas of interests include machine learning, data structures, algorithms, Natural Language Processing and artificial Intelligence. She is in teaching from 18 years in VCET, Puttur, Dakshina Kannada, Karnataka, India.