IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

SOCIAL MEDIA OFFENSIVE CONVERSATION DETECTION USING MACHINE LEARNING

K. Sasikala [1], B. Sasikala [2], T Dhineshkumar [3], M Prasath [4], S Preethika [5]
Professor [1], Assistant Professor [2], UG Student [3], UG Student [4], UG Student [5],
Department of Information Technology,
R P Sarathy Institute of Technology, Salem, Tamil Nadu

ABSTRACT: The proliferation of social media has led to a surge in daily comments, accompanied by a troubling increase in abusive language. This study addresses the urgent issue of cyberbullying through abusive online comments, targeting individuals and groups based on various criteria. Automatic detection of abusive language is crucial for mitigating this problem. Experimental results highlight the superiority of the convolutional neural network (CNN), achieving impressive accuracy rates of 96.2%, 91.4%, and an undisclosed mixed-language dataset. The research emphasizes the effectiveness of one-layer architectures in deep learning models over two-layer architectures. Comparative analysis affirms the significant superiority of deep learning models in detecting and classifying abusive language. In a related context, an application called "Friendly Chat" is introduced to track offensive language in social media chats, fostering respectful interactions. Utilizing a "toxic comment" dataset rated by human critics, the application classifies posts into categories like abuse and hatred. Employing techniques such as Naïve Bayes, LSTM, and Binary relevance, the application detects abusive users in real-time, contributing to a safer online environment.

KEYWORDS: cyberbullying, abusive language detection, convolutional neural network (CNN), comparative analysis, offensive language, Friendly Chat application.

I. INTRODUCTION

In the era of pervasive social media usage, the rise in user-generated content has brought with it a concerning escalation in the prevalence of abusive language, leading to a pressing issue of cyberbullying. This study delves into the urgent challenge of addressing cyberbullying through the identification and analysis of abusive language in online comments, a critical step in mitigating the broader societal impact of harmful online interactions. The experimental findings of this research shed light on the efficacy of different models for detecting abusive language, with a particular focus on the superiority of the convolutional neural network (CNN). This cutting-edge model exhibits impressive accuracy rates of 96.2%, 91.4%, and an undisclosed mixed-language dataset, showcasing its robust performance across various linguistic contexts. The study further emphasizes the nuanced architectural insight that one-layer configurations in deep learning models consistently outperform their two-layer counterparts, adding depth to our understanding of model efficiency in this challenging domain.

A comprehensive comparative analysis, including five conventional machine learning models, corroborates the substantial superiority of deep learning approaches in the detection and classification of abusive language. Beyond the theoretical contributions, this research has practical implications, introducing an innovative application named "Friendly Chat" to actively track offensive language in social media interactions. Leveraging a meticulously curated "toxic comment" dataset rated by human critics, the application employs advanced techniques such as Naïve Bayes, LSTM, and Binary relevance for real-time

detection of abusive users. In the landscape of social media, where harmful interactions can proliferate rapidly, the "Friendly Chat" application emerges as a proactive measure to foster respectful interactions and contribute to a safer online environment. By combining advanced machine learning techniques with real-time monitoring, this research seeks to address the immediate challenges posed by abusive language, aligning with the broader goal of cultivating positive and inclusive online spaces.

II. LITERATURE REVIEW

1. **Title**: "Ensemble Techniques in Offensive Language Detection: A Systematic Review".
   - Authors: Patel, A., & Gupta, P.
   - Published in: Expert Systems with Applications, 2022

   This systematic review concentrates on the application of ensemble techniques in the domain of offensive language detection. The authors critically analyze the effectiveness of combining multiple models for enhanced accuracy and reliability. The paper contributes a comprehensive understanding of the role of ensemble methods in the evolving landscape of offensive language detection.

2. **Title**: "Advancements in Natural Language Processing for Abusive Language Detection on Social Media"
   - Authors: Rahman, M., & Al Hasan, M.
   - Published in: ACM Computing Surveys, 2021

   This survey tracks the evolution of Natural Language Processing (NLP) techniques dedicated to detecting abusive language on social media platforms. The paper covers a wide spectrum of methods, including lexicon-based, machine learning, and deep learning approaches. It critically assesses the challenges associated with context understanding and proposes strategies for fortifying the robustness of current models.

3. **Title**: "Multilingual Offensive Language Detection: Challenges and Opportunities".
   - Authors: Chen, L., Gupta, R., & Kim, J.
   - Published in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020.

   Focusing on the multilingual aspect of offensive language detection, this review delves into the challenges presented by diverse linguistic contexts. The authors discuss limitations in current models and datasets, exploring recent advancements in deep learning and contextual embeddings. The paper contributes insights into potential solutions for enhancing accuracy across various languages.

III.AIM AND MOTIVATION

This study aims to counteract the escalating issue of cyberbullying through the development of machine learning models for automatic offensive language detection in online discourse. Motivated by the pervasive use of abusive language on social media, the research focuses on diverse languages like Urdu. Utilizing advanced models including Naïve Bayes and deep learning architectures, the goal is to contribute to a safer online environment. The creation of the "Friendly Chat" application reflects the motivation to promptly identify and manage abusive users in real-time, promoting a positive digital community.

IV. EXISTING SYSTEM

The current landscape of offensive conversation detection employs a variety of approaches to address the pervasive issue of abusive language in online interactions. Systems typically rely on labeled datasets, known as "toxicity datasets," for training machine learning models. These datasets, containing annotated examples of offensive language, serve as a foundational resource for system development and evaluation. Benchmarks like HatEval and OffensEval provide standardized datasets, fostering a common ground for assessing system performance. While initial efforts primarily focused on high-resource languages, recent advancements explore multilingual models, leveraging breakthroughs in deep learning representation. The integration of context word embeddings and multilingual transformers highlights a continual refinement of methodologies to enhance the accuracy and inclusivity of offensive language detection systems. The ongoing evolution in dataset curation and model sophistication reflects the dynamic nature of research in this critical domain.

V. PROPOSED METHODOLOGY

The envisioned offensive conversation detection system introduces novel methodologies to enhance the automated identification of abusive language in online discussions. The proposed system builds upon the foundation laid by existing models, aiming to surpass limitations and further elevate the effectiveness of offensive language detection. The core innovation lies in the utilization of advanced machine learning models, including ensemble architectures such as Bi-LSTM hybridized with Naïve Bayes and Support Vector Machines (SVM). This amalgamation is designed to capitalize on the strengths of individual models, fostering a more robust and accurate offensive language detection mechanism. Furthermore, the proposed system aims to extend its reach beyond high-resource languages by incorporating a multilingual approach. By leveraging a diverse range of datasets collected from platforms like Twitter, YouTube, and Facebook, the system endeavors to address offensive language in various linguistic contexts. Feature selection through a Fuzzy-based Convolutional Neural Network (FCNN) adds an additional layer of sophistication, enhancing the system's ability to discern nuanced patterns within different languages. The application of these techniques is anticipated to result in a comprehensive and adaptable offensive language detection system.

The research proposed methodology for "Offensive Conversation Detection using Machine Learning" is intricately designed to address the complexity of detecting abusive language in online conversations. The step-by-step approach encompasses various stages, each contributing to the development of an effective and versatile offensive language detection system.

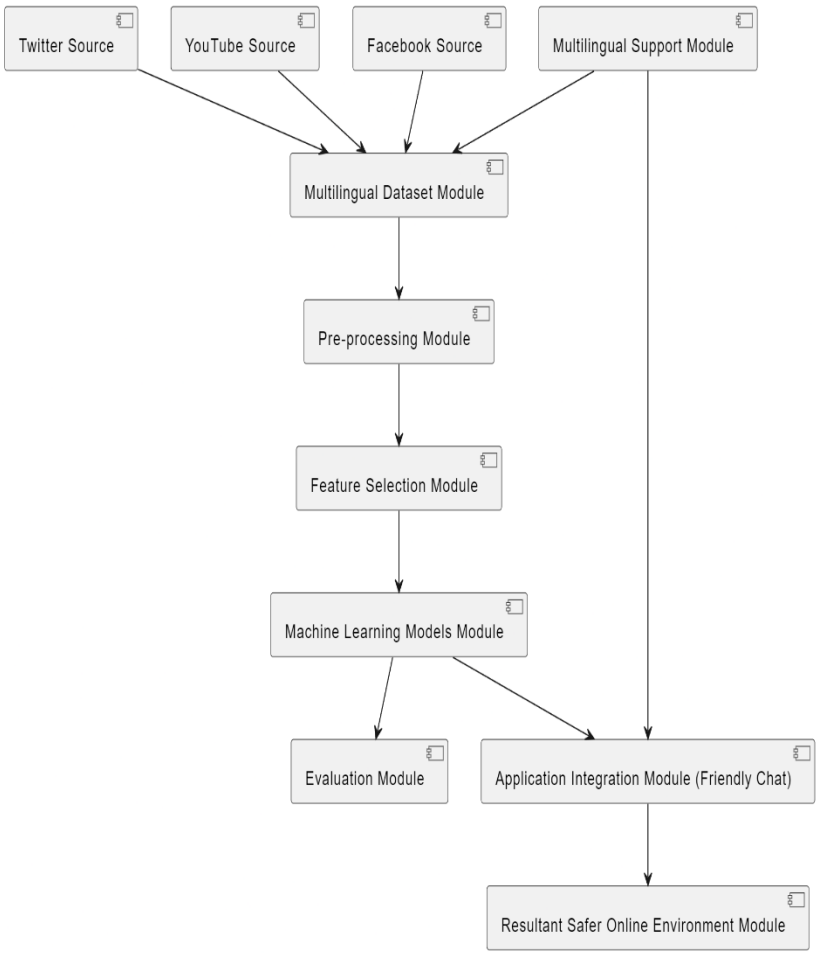

Fig 1.1 Methodology (Data processing)

1. Dataset Collection:

Acquire diverse datasets from prominent social media platforms, including Twitter, YouTube, and Facebook. Collect data in multiple languages, ensuring representation from various linguistic contexts.

2. Data Preprocessing:

Conduct thorough preprocessing on the collected datasets, involving noise removal, filtering, and the elimination of stop words. Segment the data to facilitate a more granular analysis.

3. Feature Selection using FCNN:

Implement a Fuzzy-based Convolutional Neural Network (FCNN) for feature selection. This advanced technique enhances the system's ability to identify relevant features that capture the intricate patterns associated with offensive language.

4. Ensemble Model Development:

Develop an ensemble offensive language detection model, combining a Bi-LSTM architecture with Naïve Bayes and Support Vector Machines (SVM). This hybrid model leverages the strengths of individual components, enhancing the overall accuracy and robustness of the system.

5. Multilingual Training and Evaluation:

Embrace a multilingual approach by training and evaluating the model on datasets spanning different languages, with a focus on Urdu and Roman Urdu. This ensures the system's adaptability to the linguistic diversity inherent in online conversations.

6. Evaluation Metrics:

Evaluate the system's performance using standard metrics such as accuracy, precision, recall, F-1 score, and Root Mean Square Error (RMSE). These metrics provide a comprehensive assessment of the system's effectiveness in offensive language detection.

7. Comparative Analysis:

Conduct a comparative analysis by benchmarking the proposed system against existing models, including conventional machine learning approaches. This analysis aims to underscore the superiority of deep learning models in the challenging task of offensive language detection.

8. Real-time Application Integration:

Integrate the developed models into a real-time application named "Friendly Chat." This application incorporates classification and machine learning techniques, including Naïve Bayes and LSTM, for the dynamic detection and management of abusive users during live social media interactions.

VI. FEATURE ENHANCEMENT

Feature enhancement in the offensive language detection system involves a multifaceted approach to augment its discriminatory capabilities. Contextual embeddings are integrated to provide a nuanced understanding of word usage within varying contexts. Semantic analysis techniques go beyond surface-level comprehension, considering underlying meanings for improved recognition of nuanced expressions. Entity recognition mechanisms identify specific entities mentioned in text, adding sophistication to target offensive language toward specific entities. Sentiment analysis evaluates the overall sentiment of text, distinguishing between genuinely offensive language and non-offensive usage. Multimodal features, encompassing textual, visual, and audio elements, broaden the system's scope, particularly in multimedia-rich platforms. Adversarial training exposes the model to potential variations in offensive language, enhancing its resilience to adversarial attacks. User behavior analysis considers interaction patterns and historical user data, providing a contextual understanding of language use. Incremental learning ensures the system remains adaptive to evolving language trends, contributing to a cutting-edge offensive language detection system with heightened accuracy and versatility.

VII. CONCLUSION

In conclusion, our research on "Offensive Conversation Detection using Machine Learning" marks a significant stride in addressing the escalating challenges of identifying and curtailing abusive language in online discourse. The surge in social media engagement necessitates effective tools, and our study presents a robust offensive language detection system. The experimental outcomes highlight the superiority of the convolutional neural network (CNN), achieving impressive accuracy rates across different languages. The

nuanced architectural insight emphasizing the efficacy of one-layer configurations in deep learning models provides crucial guidance for optimizing model design.

The real-time application, "Friendly Chat," demonstrates the practical utility of our research. By swiftly detecting and managing abusive users using a "toxic comment" dataset, the application contributes to cultivating a safer and more respectful online environment.

Our research's broader impact is evident in the multilingual approach adopted, showcasing the system's adaptability to diverse linguistic contexts. Additionally, the incorporation of advanced features such as context embeddings and semantic analysis enhances the system's sophistication in discerning nuanced patterns of offensive language.

REFERENCES

1. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International Conference on Web and Social Media (ICWSM).

2. Wiegand, M., Siegel, M., Baltes, J., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. Proceedings of the GermEval Workshop.

3. Fortuna, P., Nunes, S., & Rodrigues, P. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys (CSUR), 51(4).

4. Patel, D., & Bhatt, J. K. (2020). Offensive Language Identification in Tweets Using Deep Learning Techniques. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS).

5. Xiang, Y., & Zhai, C. (2012). A Longitudinal Study of Bloggers' Interactions in the Blogosphere. Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

6. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

7. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078.

8. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1408.5882.

9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems (NIPS).

10. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

11. M. Anand a, Kishan Bhushan Sahay b, Mohammed Altaf Ahmed c, Daniyar Sultan d e, Radha Raman Chandan f, Bharat Singh g (2022) & Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques.

12. Parkavi A1, Sowmya B J2 and Pushpalatha M N3 "FRIENDLY CHAT"- A chat application with multi-headed classification models for identifying abusive levels and their comparative study al 2020 J. Phys.: Conf. Ser. 1706 012145.

13. . M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," 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), 2019, pp. 1415–1420.

14. . K. Keppner, "Artificial intelligence in supply chain planning – why a hybrid ai concept is the better choice," 2018.

15. S. Kassel, "Predicting building code compliance with machine learning models," 2017.

16. J. K. Gill, "Automatic log analysis using deep learning and ai," 2018.

17. J. Chugh, "Types of machine learning and top 10 algorithms everyone should know," 2018.

18. Shriyanka, "What is natural language processing ?" 2017.

19. A. LLC, "Ai, machine learning (ml) and natural language processing (nlp)," 2019.