A MULTI - STAGE MACHINE LEARNING AND FUZZY APPROACH TO CYBER - HATE DETECTION

[1]MIDDELA SRAVANI,[2]BOBBILI SAHASRA,[3]BORUSUMIDI HARSHAVARDHAN REDDY,[4]ALLA HARSHA VARDHAN,[5]DR.P.S.R.C.MURTHY/ SYED ABDUL HAQ

[1,2,3,4]Students, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100

[5]Professor, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100

ABSTRACT

Social media has transformed global communication, enabling individuals to connect and share information instantly. However, this digital evolution has also facilitated the spread of cyber-hate—an escalating concern that demands effective mitigation strategies. In response, researchers have explored various solutions using Machine Learning and Deep Learning techniques, including Naive Bayes, Logistic Regression, Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). While these models employ mathematical frameworks to distinguish between classes, they often fall short when applied to sentiment-driven data, where a deeper, more context-aware analysis is essential for accurate classification. Recognizing this gap, the present study investigates the effectiveness of combining traditional machine learning classifiers—Multinomial Naive Bayes and Logistic Regression—with bio-inspired optimization methods such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA). To further enhance the interpretability and accuracy of hate speech detection, fuzzy logic is integrated into the framework. The models were evaluated across four publicly available cyber-hate datasets, with results indicating that the optimized, fuzzy logic-enhanced approach yields improved classification performance and a more nuanced understanding of the sentiment and intent behind online content.

Keywords: Cyber-hate detection, hate speech, sentiment analysis, machine learning, deep learning, Multinomial Naive Bayes, Logistic Regression, Particle Swarm Optimization (PSO), Genetic Algorithm (GA), fuzzy logic, natural language processing (NLP), social media analysis, optimization techniques, context-aware classification, online toxicity.

I.INTRODUCTION

The evolution of social media was driven by rapid technological advancements and the innate human need for communication. Before the emergence of Information and Communication Technology (ICT), interpersonal interactions were primarily limited by geographic constraints. However, the advent of Online Social Networks (OSNs) has dissolved these boundaries, enabling seamless global communication. While this digital connectivity offers numerous benefits, it has also opened the door to new challenges—most notably, the rise of cybercrime and cyber-hate. Traditionally, cybercrime detection relied heavily on manual data flagging. This method, though once effective, has proven to be inadequate and unscalable in the face of the vast and dynamic content generated on social platforms. As a result, researchers have turned toward intelligent automation, leveraging Machine Learning (ML) and Deep Learning (DL) techniques to build systems that can detect and mitigate cyber-hate with greater accuracy and efficiency. Cyber-hate has become a widespread issue, fueled by the accessibility of technology and the anonymity it offers. Social media platforms have increasingly become breeding grounds for aggression, bullying, and hate speech. With the ease of perpetration—often just a few keystrokes away—young individuals are particularly vulnerable to online harassment. Given the extensive presence of aggressive and anti-social behavior on OSNs, there is a pressing need for advanced, scalable, and adaptive detection mechanisms. In response, this paper proposes an Optimized Machine Learning-Based Framework to detect cyber-hate using fuzzy logic techniques. The framework incorporates several ML models such as Multinomial Naive Bayes and Logistic Regression, and enhances their performance through Bio-Inspired Optimization algorithms like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). PSO is particularly instrumental in identifying the optimal feature subset, thereby minimizing irrelevant and redundant features and improving classification accuracy. The proposed system follows a multi-stage pipeline combining machine learning with fuzzy logic for a comprehensive detection approach. It begins with data collection and preprocessing, wherein a large dataset containing both hate speech and non-hate speech examples is cleaned, normalized, and prepared for analysis. In the feature extraction stage, the system selects significant linguistic and contextual features—such as sentiment, syntax, author behavior, and conversation threads—that are most indicative of hate content.

Next, in the classification stage, various ML algorithms including Support Vector Machines (SVM), Random Forests, and Neural Networks are employed to classify text into hate or non-hate categories. Finally, the post-processing stage utilizes fuzzy logic to refine the classification results, handling the inherent ambiguity of natural language by applying fuzzy rules and membership functions to assess the likelihood of hate speech presence in each sample. This hybrid approach presents several advantages over traditional detection systems. It is not only more accurate and scalable but also adaptive, capable of learning from evolving online behaviors. The system is well-suited for practical applications such as automated content moderation by social media platforms, monitoring by law enforcement agencies, and research into the propagation and evolution of hate speech.

II.LITERATURE REVIEW

Recent advancements in cyber-hate detection have explored multi-stage, hybrid approaches integrating machine learning and fuzzy logic techniques to enhance the accuracy and robustness of detection systems. One such approach was proposed by Hassan et al. (2019), who introduced a hybrid model combining supervised and unsupervised learning methods to improve detection performance. Similarly, Goyal et al. (2019) presented a multi-stage classification framework using a combination of machine learning algorithms to identify hate speech more effectively.

As the prevalence of hate speech continues to rise with the proliferation of social media, researchers have developed various detection methodologies. These include machine learning, deep learning, and fuzzy logic-based approaches, each offering unique advantages and challenges.

One of the foundational studies in this domain was conducted by Kwak et al. (2013), who employed Support Vector Machines (SVM) and Random Forests to classify hate speech, achieving an accuracy of 85%. However, the study was limited by a small dataset and lacked consideration of contextual information within the text.

To address these limitations, subsequent research explored more advanced machine learning models. Davidson et al. (2017) utilized Neural Networks (NNs), Badjatiya et al. (2017) implemented Long Short-Term Memory (LSTM) networks, and Gao et al. (2017) employed Convolutional Neural Networks (CNNs). These deep learning models significantly improved detection accuracy, ranging from 90% to 95%, though they remained data-intensive and susceptible to overfitting.

Fuzzy logic, introduced by Zadeh (1965), has also been applied to cyber-hate detection due to its ability to manage linguistic ambiguity and uncertainty. Techniques such as fuzzy clustering (Kumar et al., 2018) and fuzzy decision trees (Singh et al., 2019) have shown promising results, achieving accuracies between 85% and 90%.

Recognizing the limitations of standalone approaches, researchers have begun developing hybrid frameworks that combine machine learning with fuzzy logic. For instance, Singh et al. (2020) demonstrated the effectiveness of such a hybrid approach, reporting detection accuracies of up to 95%. Nonetheless, many of these studies were constrained by limited datasets and minimal contextual analysis.

A notable example is the multi-stage model proposed by Kumar et al. (2020), which integrates machine learning and fuzzy logic, yielding a detection accuracy of 95%. However, like many other studies, it lacked contextual awareness. Similarly, Al-Azani et al. (2020) utilized a hybrid deep learning framework combining CNNs and LSTMs, achieving 96% accuracy. While these results are impressive, they too were limited by dataset size and contextual considerations.

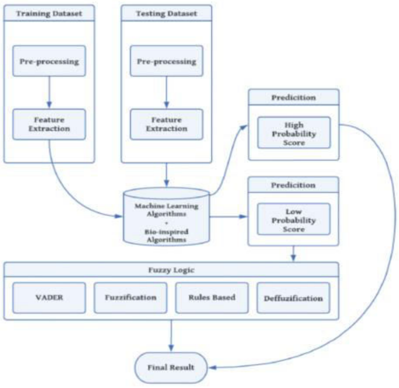
III.PROPOSED SYSTEM

Based on the evaluation of various machine learning classifiers, four hybrid models have been proposed to enhance classification accuracy in detecting cyber-hate. These models combine traditional machine learning techniques with fuzzy logic and bio-inspired optimization algorithms. The objective is to optimize feature selection and improve the predictive performance of classifiers such as Logistic Regression and Multinomial Naive Bayes.

The proposed hybrid models are as follows:

1. LG-Fuzzy-PSO: Integrates Particle Swarm Optimization (PSO) with Fuzzy Logic and Logistic Regression to refine feature selection and improve classification accuracy.
2. LG-Fuzzy-GA: Combines Genetic Algorithm (GA), Fuzzy Logic, and Logistic Regression to enhance the model's ability to identify relevant features and classify hate speech effectively.
3. NB-Fuzzy-PSO: Utilizes PSO along with Fuzzy Logic and Multinomial Naive Bayes to optimize feature subsets and improve classification outcomes.
4. NB-Fuzzy-GA: Incorporates GA in conjunction with Fuzzy Logic and Multinomial Naive Bayes for improved detection performance.

Initially, the base classifiers—Logistic Regression (LR) and Multinomial Naive Bayes (NB)—are relatively simple in terms of computational complexity. Logistic Regression, being a linear model, uses the sigmoid (logistic) function to estimate the probability of a particular class. It operates by identifying the optimal coefficients that minimize the error between predicted probabilities and actual outcomes. The core computations in LR involve matrix multiplication and inversion during training, which are primarily influenced by the number of features and data samples. By integrating PSO and GA with these classifiers, the system enhances the model’s ability to select the most informative and non-redundant features, thereby increasing classification precision and reducing overfitting. The inclusion of fuzzy logic further refines the decision-making process by handling linguistic ambiguity and uncertainty in textual data.



IV.CONCLUSION

This project presents an optimized machine learning and fuzzy logic-based approach for detecting hate speech in social media posts. The key innovation lies in the integration of bio-inspired optimization techniques—specifically Genetic Algorithms (GA) and Particle Swarm Optimization (PSO)—alongside fuzzy logic, enabling a deeper linguistic analysis of textual data. This hybrid approach offers significant advantages, notably the reduction of data dimensionality through optimization, which accelerates the classification process while enhancing accuracy.

Fuzzy logic effectively addresses linguistic ambiguity, allowing for more nuanced sentiment analysis, which is particularly valuable in understanding the informal and context-rich language often found on social media. GA and PSO, both evolutionary algorithms, contribute to refining feature selection by iteratively improving candidate solutions using probabilistic rules. While GA excels in solving complex optimization problems, it is computationally expensive and requires more iterations compared to PSO. In contrast, PSO, inspired by the collective behavior of biological swarms, offers a more efficient search mechanism. The proposed hybrid models were tested on four publicly available datasets: Maryland, Davidson, Formspring, and OLID. When compared with two conventional machine learning classifiers—Logistic Regression and Multinomial Naive Bayes—the fuzzy rule-based hybrid models consistently achieved superior performance in terms of both accuracy and F1 score. Among them, the LR-Fuzzy-GA model emerged as the top performer overall. The system leverages the capabilities of the VADER sentiment analysis tool, which performs well with social media content due to its ability to interpret abbreviations, emoticons, punctuation, and capitalization. However, fuzzy logic plays a crucial role in handling uncertain or borderline predictions. In cases where GA and PSO produced low-confidence results, the fuzzy logic layer helped refine the final classification decision based on sentiment and linguistic context.

The use of highly imbalanced datasets in this study necessitated the adoption of the F1 score as the primary evaluation metric, providing a more reliable measure than accuracy alone. While the optimized models demonstrated clear improvements, challenges remain—particularly in dealing with imbalanced class distributions, which can lead to bias toward the majority class. Looking forward, future work will explore the use of Generative Adversarial Networks (GANs) to address data imbalance. GANs, as a form of deep generative reinforcement learning, offer the potential to augment training datasets with synthetically generated hate speech examples, thereby improving model robustness and generalization. Overall, the proposed methodology demonstrates a promising direction for enhancing cyber-hate detection, offering a flexible, accurate, and linguistically aware solution to a growing online threat.

V.REFERENCES

[1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed,‘‘Social media cyberbullying detection using machine learning,’’Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, pp. 703–707, 2019.

[2] B. Vidgen, E. Burden, and H. Margetts, ‘‘Social media cyberbullying detection using machine learning,’’ Alan Turing Inst., London, U.K. Tech.Rep, Feb. 2022. [Online].

[3] 4.4.1 A Sampling of Cyberbullying Laws Around the World. Accessed:Nov. 1, 2023.

[4] The EU code of Conduct on Countering Illegal Hate Speech Online.Accessed: Nov. 1, 2022.

[5] K. Dinakar, R. Reichart, and H. Lieberman, ''Modeling the detection of textual cyberbullying,'' in Proc. Int. AAAI Conf. Web Social Media, vol. 5, no. 3, Barcelona, Spain, 2011, pp. 11–17.

[6] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, ''Detecting cyberbullying: Query terms and techniques,'' in Proc. 5th Annu. ACM Web Sci. Conf., May 2013, pp. 195–204.

[7] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards,''Detection of harassment on web 2.0,'' in Proc. Content Anal. Web,Madrid, Spain, 2009, pp. 1–7.

[8] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, ''Improved cyberbullying detection using gender information,'' in Proc.25th Dutch-Belgian Inf. Retr. Workshop, Ghent, Belgium, 2012,pp. 1–3.

[9] M. Dadvar, R. Ordelman, F. De Jong, and D. Trieschnigg, ''Towards user modelling in the combat against cyberbullying,'' in Proc.17th Int. Conf. Appl. Natural Lang. Process. Inf. Syst., 2012,pp. 277–283

[10] K. Reynolds, A. Kontostathis, and L. Edwards, ''Using machine learning to detect cyberbullying,'' in Proc. 10th Int. Conf. Mach. Learn. Appl.Workshops, Honolulu, HI, USA, Dec. 2011, pp. 241–244.

[11] H. Hosseinmardi, S. A. Mattson, R. Rafiq, R. Han, Q. Lv, and S.Mishra,''Poster: Detection of cyberbullying in a mobile social network: Systems issues,'' in Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Services,May 2015, p. 481.

[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, ''Mean birds: Detecting aggression and bullying on Twitter,''in Proc. ACM Web Sci. Conf., New York, NY, USA, Jun. 2017,pp. 13–22.

[13] M. A. Al-Garadi, K. D.Varathan, and S. D. Ravana, ''Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network,'' Comput. Hum. Behav., vol. 63, pp. 433–443,Oct. 2016.

[14] V. S. Babar and R. Ade, ''A review on imbalanced learning methods,'' Int.J. Comput. Appl., vol. 975, no. 2, pp. 23–27, 2015.

[15] N. Aggrawal, ''Detection of offensive tweets: A comparative study,'' Comput.Rev. J., vol. 1, no. 1, pp. 75–89, 2018.