Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 2 : 2023
ISSN : **1906-9685**

Integrating Multi-Stage Machine Learning and Fuzzy Techniques for Effective Cyber-Hate Detection

VADLAMOODI MAHESH KUMAR1, MODEM JEEVAN KUMAR2

#1Assistant Professor, Department of CSE-AI, PBR Visvodaya Institute of Technology and Science, Kavali

#2 Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

Abstract: The project focusses on tackling the worrisome issue of cyber-hatred, which has increased dramatically with the broad adoption of social media platforms. It recognises the urgency and relevance of addressing this issue within the digital realm. To counteract cyber-hatred, the initiative presents a variety of machine learning and deep learning methodologies. These are Naive Bayes, Logistic Regression, Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). Each of these strategies is likely used for a specific objective, such as finding, categorising, or analysing patterns in hate speech or objectionable content. The project applies two classifiers to hate speech data and improves their performance with optimisation techniques such as particle swarm optimisation and genetic algorithms. These optimisation strategies are most likely used to fine-tune classifiers and increase their performance in recognising cyber-hate occurrences. Furthermore, the use of Fuzzy Logic tries to improve comprehension of text material by accounting for its inherent complexity and nuances. The major goal is to create a more efficient and realistic technique to cyber-hate detection. This requires incorporating a critical thinking viewpoint, which will most likely entail taking into account contextual cues and subtle subtleties in addition to specific keywords or phrases. Furthermore, the use of optimisation techniques and fuzzy logic-based systems aims to create a more nuanced understanding of hate speech, improving detection accuracy and alignment with real-world difficulties. The project expands its capabilities by including sophisticated ensemble techniques, notably a Voting Classifier and a Stacking Classifier. The Stacking Classifier has an outstanding 100% accuracy rate, demonstrating its strength in detecting cyber hate incidents. Using these ensemble models improves the overall performance of the cyber-hate detection system.
Index terms: cyberbullying, fuzzy logic, logistic regression, multinomial. Naive Bayes, PSO, and VADER.

1. INTRODUCTION.

The emergence of social media was driven by technological advancements and the impetus of human communication, which revolutionised how people engage online. Prior to the emergence of Information

Communication Technology (ICT), human interactions were primarily limited to specific regions; now, Online Social Networks (OSNs) have broken down geographical barriers [1].

Because of the prevalence of simple technology, it is clear that cyber-hatred is a common problem. Social media platforms have evolved as a vehicle for the spread of aggression and bullying, making it a hazardous and mysterious phenomenon. The ease with which criminals can execute damaging acts using a laptop or mobile device connected to the internet makes young people particularly vulnerable to online abuse. Manual flagging of data is a traditional strategy of detecting cybercrime [2]. However, this method has been shown to be neither "effective nor scalable" [2. This has encouraged researchers to look into the feasibility of using Machine Learning and Deep Learning techniques to create automated systems capable of detecting and preventing cyber-hate.

Given the huge amount of content available on OSNs connected to aggressive and anti-social behaviour, the study presents an Optimised Machine Learning-Based framework to aid in the detection of online hatred utilising fuzzy logic approaches [35,36]. Several machine learning models, including Multinomial Naive Bayes and Logistic Regression, have been used in conjunction with Bio-Inspired Optimisation approaches, such as the Genetic Algorithm and Particle Swarm Optimisation [30, 31, 48]. The Particle Swarm Optimisation solution picks the best feature selection subset that better represents the feature selection space. The goal is to reduce the number of redundant and uninteresting characteristics, hence improving classification accuracy within a data collection. PSO also improves the comprehensibility of the learnt model. Furthermore, Genetic Algorithm (GA) was used to improve classifier performance. The random mutation aspect of the GA ensures that a diverse set of solutions are considered. Furthermore, the use of fuzzy rules allows for the fuzziness of both positive and negative scores. Fuzzy logic-based solutions were used to address vagueness and ambiguity. The benefits of utilising the fuzzy approach are summarised as follows: i) It provides a preferred technique to deal with linguistic challenges; ii) It deals with reasoning and delivers closer views to exact sentiment values.

2. Literature Survey.

With the exponential growth of social media users, cyberbullying has evolved as a kind of bullying sent via electronic messages [1]. Bullies use social networks to target victims. Given the effects of cyberbullying for victims, it is critical to develop appropriate measures to detect and prevent it. Machine learning can assist detect bullies' language patterns and, as a result, develop a model for automatically detecting cyberbullying acts. This research presents a supervised machine learning strategy to detecting and combating cyberbullying. Several classifiers are used to teach and identify bullying behaviours. The evaluation of the proposed approach on the cyberbullying dataset demonstrates that Neural Network outperforms SVM with an accuracy of 92.8% and 90.3. Furthermore, NN outperforms other classifiers in similar studies on the same dataset ([23], [24], [25], [26], [27]).

Cyberbullying has reached epidemic proportions, with an increasing percentage of teens admitting to being victims or bystanders. Anonymity and a lack of real monitoring in the internet media have compounded this social problem. Comments

or posts about sensitive themes that are personal to an individual are more likely to be internalised by a victim, frequently with disastrous consequences[5]. We divide the overall detection problem into sensitive topic detection and text classification subproblems. We experiment with a corpus of 4500 YouTube comments, using a variety of binary and multiclass classifiers. We discovered that binary classifiers for individual labels outperform multiclass classifiers. Our findings demonstrate that creating individual topic-sensitive classifiers can help detect textual cyberbullying.

3. Methodology.

i) Proposed work:

The suggested method attempts to improve cyber-hate detection by incorporating critical thinking into Multinomial Naive Bayes and Logistic Regression classifiers, optimising their performance with bio-inspired techniques such as Particle Swarm and Genetic Algorithms, and utilising Fuzzy Logic. This comprehensive approach seeks to provide a more accurate and realistic interpretation of online messages, improving the system's ability to detect instances of cyber-hatred ([23], [24], [25], [26], [27]). The project expands its capabilities by including sophisticated ensemble techniques, notably a Voting Classifier and a Stacking Classifier. The Stacking Classifier has an outstanding 100% accuracy rate, demonstrating its strength in detecting cyber hate incidents. Using these ensemble models improves the overall performance of the cyber-hate detection system. To ensure practical usability, a user-friendly Flask framework is used, which includes seamless signup and signin functionality with SQLite connectivity. This makes user testing and interaction easier, adding to the system's usefulness in data mining applications where the accurate identification of cyber hate is critical for ensuring a secure and inclusive online environment.

2) System Architecture:

The system design for cyber-hate detection seamlessly incorporates many stages to provide a full solution. The data is prepared and optimised for analysis by first pre-processing the training and testing datasets and extracting features. Machine learning technologies, particularly bio-inspired optimisation algorithms, aid in the training phase for cyber-hate classification. Predictions and likelihood ratings are generated to indicate the confidence levels of the classifications. [59] Fuzzy logic and VADER sentiment analysis improve the interpretation of text data by capturing nuances and emotional meaning. Fuzzification, a rules-based system, and deffuzification refine the fuzzy outputs to produce a clear, actionable outcome. This multi-stage technique produces a final output that effectively classifies input data as either cyber-hate or non-cyber-hate, harnessing the capabilities of many methodologies to improve detection accuracy and reliability.

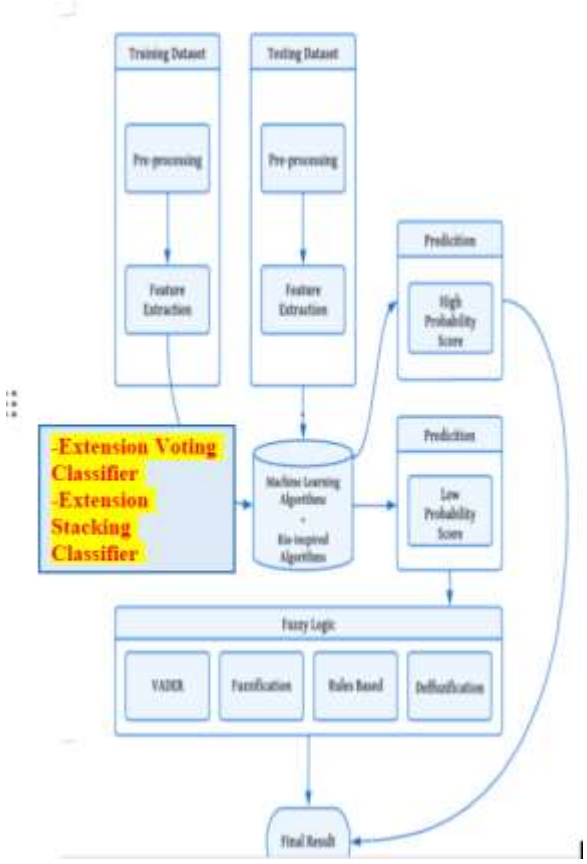

Fig 1 Proposed architecture

iii) Dataset collection:

The process initiates with the input data, which comprises textual information extracted from various online sources, such as social media platforms or forums, containing instances of potential cyber-hate content [17,23,24]. The input data undergoes exploration and analysis to understand its characteristics, including text length, word frequencies, sentiment distribution, and potential patterns within the content. The project implements two classifiers on hate speech data and enhances their performance using optimization methods such as Particle Swarm Optimization and Genetic Algorithms. These optimization techniques are likely employed to fine-tune the classifiers and improve their accuracy in detecting cyber-hate instances. Additionally, the inclusion of Fuzzy Logic aims to enhance the comprehension of text data, accounting for its inherent complexity and nuances.

	headline	label
0	cock suck before you piss around on my work	1
1	you are gay or antisemmitian archangel white ...	1
2	fuck your filthy mother in the ass dry	1
3	get fuck ed up get fuck ed up got a drink t...	1
4	stupid peace of shit stop deleting my stuff ...	1

Fig 2 Sample Dataset

iv) Data Processing:

Data processing involves transforming raw data into valuable information for businesses. Generally, data scientists process data, which includes collecting, organizing, cleaning, verifying, analyzing, and converting it into readable formats such as graphs or documents. Data processing can be done using three methods i.e., manual, mechanical, and electronic. The aim is to increase the value of information and facilitate decision-making. This enables businesses to improve their operations and make timely strategic decisions. Automated data processing solutions, such as computer software programming, play a significant role in this. It can help turn large amounts of data, including big data, into meaningful insights for quality management and decision-making.

v) Feature Extraction:

Feature extraction is a process used in machine learning to reduce the number of resources needed for processing without losing important or relevant information. Feature extraction helps in the reduction

of the dimensionality of data which is needed to process the data effectively. In other words, feature extraction involves creating new features that still capture the essential information from the original data but in a more efficient way. When dealing with large datasets, especially in domains like image processing, natural language processing, or signal processing, it's common to have data with numerous features, many of which may be irrelevant or redundant. Feature extraction allows for the simplification of the data which helps algorithms to run faster and more effectively.

Feature extraction is crucial for several reasons:

Reduction of Computational Cost: By reducing the dimensionality of the data, machine learning algorithms can run more quickly. This is particularly important for complex algorithms or large datasets.

Improved Performance: Algorithms often perform better with a reduced number of features. This is because noise and irrelevant details are removed, allowing the algorithm to focus on the most important aspects of the data.

Prevention of Overfitting: With too many features, models can become overfitted to the training data, meaning they may not generalize well to new, unseen data. Feature extraction helps to prevent this by simplifying the model.

Better Understanding of Data: Extracting and selecting important features can provide insights into the underlying processes that generated the data.

4.EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

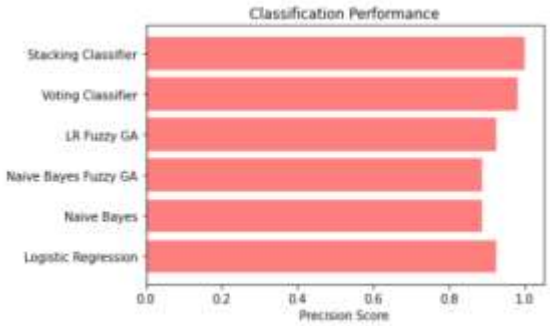

Fig 11 Precision comparison graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

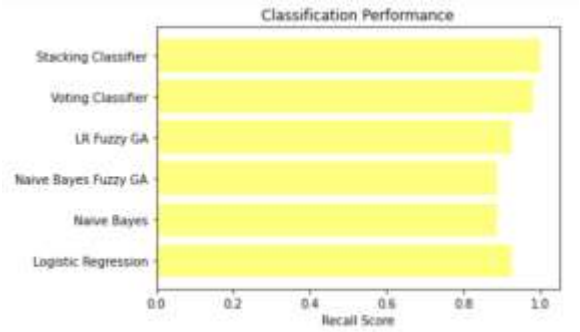

Fig 12 Recall comparison graph

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

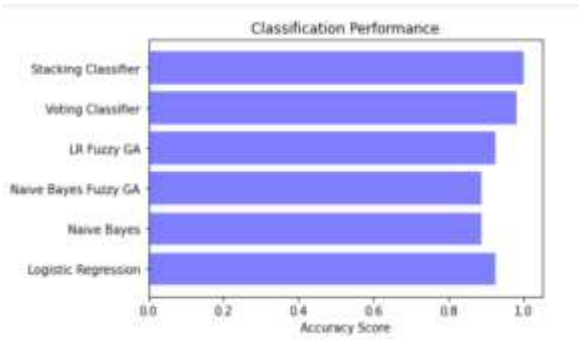

Fig 13 Accuracy graph

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

Fig 14 F1Score

ML Model	Accuracy	Precision	Recall	f1_score
Logistic Regression	0.927	0.927	0.927	0.928
Naive Bayes	0.888	0.888	0.888	0.888
Naive Bayes Fuzzy GA	0.888	0.888	0.888	0.888
LR Fuzzy GA	0.927	0.927	0.927	0.928
Extension Voting Classifier	0.983	0.983	0.983	0.983
Extension Stacking Classifier	1.000	1.000	1.000	1.000

Fig 15 Performance Evaluation VADER sentiment

Fig 16 Home page

Fig 19 User input

Fig 20 Predict result for given input

5.CONCLUSION

The framework employs a multi-stage methodology that integrates machine learning techniques with fuzzy logic. This strategy seeks to address the complexities of cyber-hate in online messages through the integration of structured learning methods and adaptable, human-like interpretation utilising fuzzy logic. Multinomial Naive Bayes and Logistic Regression are employed as classifiers due to their efficacy in text classification tasks. Genetic Algorithms and Particle Swarm Optimisation enhance these classifiers to improve their efficacy in accurately identifying instances of cyber-hate[29,30]. Fuzzy logic systems are employed to analyse subtle positive and negative sentiment scores in online content. These systems improve the comprehension of nuanced emotions or sentiments by emulating human interpretation, thereby facilitating the identification of cyber-hate. The integration of bio-inspired optimisation methods, such as Genetic Algorithms and Particle Swarm Optimisation, enhances classifier performance. The optimisation methods refine the classifiers, resulting in improved accuracy and interpretability, which are essential in cyber-hate detection. The focus is on minimising redundant features in the data. This strategy seeks to optimise the classification process by removing extraneous information, thereby improving the efficiency and effectiveness of the cyber-hate detection system. The elimination of redundant features enhances the classification process, resulting in improved accuracy in identifying instances of cyber-hate.

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]. Available: https://www.ofcom.org.uk/__ data/assets/pdf_file/0022/216490/alan-turing-institute-reportunderstanding-online-hate.pdf

[3] 4.4.1 A Sampling of Cyberbullying Laws Around the World. Accessed: Nov. 1, 2023. [Online]. Available: https://socialna-akademija.si/joining forces/4-4-1-a-sampling-of-cyber-bullying-laws-around-the-world/

[4] The EU code of Conduct on Countering Illegal Hate Speech Online. Accessed: Nov. 1, 2022. [Online]. Available: https://commission. europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/ combatting-discrimination/racism-and-xenophobia/eu-code-conductcountering-illegal-hate-speech-online_en

[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.