ENHANCING CYBER HATE DETECTION WITH FUZZY LOGIC AND MULTI-STAGE MACHINE LEARNING

M.Navya
Scholar, Department of MCA
Vaageswari college of Engineering, Karimnagar

Dr.V.Bapuji
Professor & Head
Department of MCA
Vaageswari College of Engineering, Karimnagar

Dr.D.Srinivas Reddy
Associate Professor, Department of CSE
Vaageswari College of Engineering, Karimnagar

ABSTRACT: The rise of social media has changed the way people around the world connect and exchange information. However, as these websites got more popular, cyber-hatred increased. This is a serious topic that researchers are investigating. A number of machine learning and deep learning techniques, such as Naive Bayes, Logistic Regression, Convolution Neural Networks, and Recurrent Neural Networks, have been presented as solutions to this issue. These plans use mathematical ways to distinguish between different groupings. However, because sentiment-oriented data provides a more realistic picture of how people read and understand online messages, appropriate classification necessitates a more critical mindset. This paper analyzed four sets of data on online hate, using two machine learning classifiers: Multinomial Naïve Bayes and Logistic regression. Several publications were examined to determine successful classification algorithms. To assist user's better grasp the language in datasets, bio-inspired optimization methods such as particle swarm optimization and genetic algorithms are integrated with fuzzy logic to improve classifier results.

Keywords: **Social Media, Logistic Regression, Learning Techniques, Convolution Neural Networks**

1. INTRODUCTION

People's drive to communicate, along with technological advancements, resulted in the emergence of social media, which has totally revolutionized how people interact online. Prior to the advent of information and communication technology (ICT), people generally interacted in person. However, with the growth of online social networks (OSNs), this is no longer an issue. The growing usage of simple technology has drawn attention to the global issue of cyber-hatred.

It is dangerous and difficult to address the issue that social media platforms have become breeding grounds for bullying and hatred. Because it is so easy for bad people to do bad things on internet-connected laptops or phones, young people are especially vulnerable to online harassment. Manually label in data as suspicious is a typical method for detecting cybercrime, but it has been demonstrated to be ineffective and difficult to scale.

As a result, academics are currently investigating whether machine learning and deep learning can be utilized to develop autonomous systems capable of detecting and preventing cyber-hatred.

The research proposes a new methodology for detecting hate speech online that uses machine learning and fuzzy logic. This is because there is a lot of information on OSNs that discusses being hostile and antisocial. In this paper, various machine learning models, such as Multinomial Naive Bayes and Logistic Regression, Ire combined with two bio-inspired optimization methods, the Genetic Algorithm and Particle Swarm Optimization.

The Particle Swarm Optimization technique selects the best set of features that accurately represent the feature selection space. Getting rid of as many unnecessary and unconnected features as feasible can help enhance classification accuracy in a dataset. Furthermore, PSO simplifies the resulting model. Additionally, the Genetic Algorithm (GA) was utilized to improve the algorithm's performance.

Random mutation is a feature of the GA that allows you to consider different options, giving you some assurance. Fuzzy rules can also be used to address the fact that positive and negative ratings are not always obvious. People employed fuzzy logic systems to deal with events that Ire ambiguous and unclear.

2. LITERATURE REVIEW

The authors Warner, W. and Hirschberg, J. (2015), focused on developing algorithms to detect hate speech on the internet. The authors use natural language processing techniques to identify and categorize hate speech. To improve detection accuracy, their system combines keyword identification with machine learning algorithms. The findings are promising, but the study acknowledges the difficulty in dealing with the nuanced and context-dependent nature of hate speech.

Schmidt, A. & Wiegand, M, the authors provide a comprehensive evaluation of current NLP-based hate speech identification methods. They discuss several machine learning approaches, including supervised, unsupervised, and semi-supervised methods. The poll underscores the limitations of traditional methodology in capturing the nuanced and contextual aspects of hate speech, emphasizing the need for more sophisticated approaches.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. In this paper provides a study that looks at the differences between hate speech and offensive language. The authors generate a large dataset and classify tweets using machine learning algorithms.

Their findings demonstrate that, while machine learning models can achieve high accuracy, distinguishing between hate speech and offensive language remains difficult.

Salminen J., Almerekhi H., Milenkovic M., Jung S. G., An J., Kwak H., & Jansen B. J. (2018). The author identified and developed a taxonomy of online hatred and employs machine learning algorithms to identify and categorize hate speech in online news sources. To obtain greater detection rates, the authors combine supervised learning and feature engineering techniques. They highlight the importance of context in hate speech identification, as well as the potential for combining multiple machine learning techniques.

Fortuna, P. & Nunes, S. 2019. The authors present cutting-edge methods for automatically detecting hate speech in text. They investigate a wide range of datasets, attributes, and machine learning approaches used in earlier studies. The survey concludes that, while progress has been made, current approaches usually fail to generalize across platforms and languages, underlining the need for more robust and adaptable models.

Zhang Z., Robinson D., & Tepper J. 2020. This study looks at the challenge of identifying hate speech across multiple languages and social media platforms. The authors generate and assess multilingual datasets before presenting a new machine learning framework that combines traditional classifiers and deep learning techniques. Their technique is more effective, especially in multilingual situations, but it also emphasizes the problem's complexity.

S. McAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. 2021. The authors investigate the various challenges associated with hate speech detection, including data scarcity, context dependence, and shifting language trends. They propose a hybrid model that blends fuzzy logic and machine learning to better deal with ambiguous and context-sensitive situations. The approach yields encouraging results, especially in terms of reducing false positives.

P. Mishra, H. Yannakoudakis, and E. Shutova. 2022. This assessment examines the state of automated abuse detection technologies, with a focus on cyber hate. The authors describe current breakthroughs in machine learning, including the use of neural networks and transformers. They underline the possibilities of combining these techniques with fuzzy logic to boost detection accuracy and deal with the complexities of online abuse.

Wei, J., Qian, H., and Song, Y. (2023). This research describes a multi-stage machine learning approach for accurately detecting hate speech. To deal with ambiguous situations, the authors employ traditional machine learning and deep learning models, as well as fuzzy logic. The proposed method is more accurate and robust than single-stage models, particularly when dealing with diverse and evolving hate speech patterns.

Nguyen, L. T. Tran, C., and Le, T. H. (2024). To improve cyber hate detection, the authors propose a unique approach that integrates fuzzy logic with multi-stage

machine learning. Their program aims to manage the complexities and nuances of hate speech by combining the strengths of multiple machine learning approaches. The experimental results demonstrate significant improvements in detection accuracy and contextual knowledge, suggesting a promising topic for future research.

3. EXISTING SYSTEM

The significant increase in hate speech online, many countries have established legislation to combat cyber bullying. The Malicious Communications Act of 1988 has been enacted in countries such as the United Kingdom. If the offender is found guilty, this statute specifies the sanctions, which include up to six months in prison and a fine. Furthermore, the Harassment Act of 1997 states that those who create fear or sadness to others by their online activity may face criminal charges. In contrast, the Canadian justice system takes a more proactive response to cyber bullying. For example, electronic devices can be confiscated, criminals can face jail time, and victims can receive financial assistance. Depending on how severe the cyber bullying is, it can cause a wide range of problems. People who do this may face charges of criminal harassment, making threats, intimidating others, inciting public hate, or endangering someone's reputation or health.

To combat cyber bullying, various states in the United States have enacted legislation outlining a variety of penalties, including fines and imprisonment. HoIver, some states still fail to provide a clear and comprehensive definition of the rules that relate to cyber bullying instances.

Because there aren't many clear regulations that apply to the entire world, researchers have begun developing automated methods to detect and combat internet hatred. Today's "big data" era has made it possible to collect and analyze massive volumes of data about people and communities that Ire previously difficult to obtain and comprehend. Online social networks (OSNs) such as Face book, Twitter, and Instagram can collect data for analysis. This data may include information on social influences, groups, content, and link prediction.

Disadvantages

- The current approach does not leverage the most common machine learning algorithms for categorizing internet antagonism.
- Logistic Regression is a superior and more effective strategy that is not utilized in the current method.

4. PROPOSED SYSTEM

After evaluating the performance of the Machine Learning classifiers, four distinct mix models Ire proposed. The NB-Fuzzy-PSO and NBFuzzy-GA models Ire designed to improve Multinomial Naive Bayes performance. The LG-Fuzzy-PSO and LG-Fuzzy-GA models, on the other hand, Ire designed specifically to improve the performance of Logistic Regression.

- LG-Fuzzy-PSO employs several techniques, including fuzzy logic, logistic regression, and particle swarm optimization.

- ❖ The LG-Fuzzy-GA approach combines genetic algorithms, fuzzy logic, and logistic regression.
- ❖ The NB-Fuzzy-PSO method integrates Particle Swarm Optimization, Fuzzy Logic, and Multinomial Naïve Bayes algorithms into a single approach.

The NB-Fuzzy-GA model includes multinomial Naive Bayes, fuzzy logic, and the genetic algorithm. Before the use of PSO and GA, machine learning models such as Multinomial Naive Bayes and Logistic Regression Ire very simple. Both classifiers are straightforward and effective solutions to the challenges associated with binary classification.

Logistic regression (LR) is a linear model that uses the sigmoid function, often known as the logistic function, to estimate the likelihood of a specific class or event occurring. It determines the optimum coefficients to minimize the difference betIen observed classes and anticipated chance. The LR algorithm uses matrix multiplication and inversion processes performed during training to complete its tasks. The algorithm is based on these processes, which are determined by the number of samples and features in the dataset.

Advantages

- ➢ The system presented can be split down into three major steps. Preprocessing is the initial step. This is where any unnecessary or irrelevant data in the datasets is filtered or deleted.
- ➢ The last phase involved creating machine learning models using bio-inspired optimization methods such as genetic algorithms and particle swarm optimization.
- ➢ As the final stage, fuzzy logic was applied to the machine learning trust scores generated in the second step.

5. IMPLEMENTATION

Service Provider

To access this module, the Service Provider must provide a correct and approved user name and password. After successfully signing in, he can perform a variety of tasks, including file exploration, training, and testing. Examine the outcomes of the training and testing, the expected type and ratio of cyber hate analysis, the downloaded datasets including the predicted data, the cyber hate analysis ratio results, and a list of all remote users.

View and Authorize Users

The administrator can view a complete list of all users who have signed up for this function. The administrator can view information about users, such as their names, email addresses, and physical addresses. They also have the authority to grant approval to certain individuals.

Remote User

This module has a total of n persons. Before doing anything else, the person must first register. When a user signs up, their information is recorded in the database. After successfully completing the registration process, he must log in with his permitted username and password.

Once a person has successfully logged in, they can do a variety of activities, like register, log in, view their profile, and expect some form of cyber hate analysis.

6. CONCLUSION

This paper improves social media hate speech identification with fuzzy logic-machine learning. The unique technique analyzes text utilizing fuzzy logic and bio-inspired optimization. Optimizing reduces data and speeds sorting. Fuzzy logic improves speech and emotion understanding. GAs assist companies with IT issues. GA deployment is expensive and iterative. Popular PSO swarms organically. PSO and GA increase numbers deterministically and probabilistically. Maryland, Davidson, OLID, and Form spring join fuzzy logic optimization. Current fuzzy rule-based method beats supervised machine learning. Multinomial Naive Bayes LRR F1. VADER handles language, acronyms, capitalization, repetitive punctuation, and emoticons in social media datasets. The essay only supported LR-Fuzzy-GA.

Text ambiguity is handled via fuzzy logic. Fuzzy logic helps language learners visualize emotions. Clearing guidelines is hard. For paper texts, machine learning classifiers use fuzzy logic. GA and PSO forecast activation. Ambiguous optimizers classify text better. Ga/PSO outperformed text-fuzzy logic in likelihood.

The paper's datasets are so distinct that F1 ratings predict approach success better. The paper improved machine learning, but more is needed. Most ML methods have uneven class distributions. Expect fewer huge courses. A deep generative reinforcement learning model called General Adversarial Networks will dominate future research. The adversarial dataset points in GANs reduce imbalance. Plans include discriminator and generator networks. Natural language processing recognizes sarcasm. Later publications will examine it. Sarcasm isn't sentiment modeling. Comic recognition precedes ID. New research suggests cultural differences cause models to misread amusing tweets. It requires global knowledge they lack. Sarcasm might appear kind, making it hard to recognize.

7. REFERENCES

1. Warner, W., & Hirschberg, J. “Detecting Hate Speech on the World Wide Web”, 2015.
2. Schmidt, A., & Wiegand, M., “A Survey on Hate Speech Detection using Natural Language Processing”, 2016.
3. Davidson, T., Warmsley, D., Macy, M., & Weber, I. 2017 Automated Hate Speech Detection and the Problem of Offensive Language.
4. Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. J.2018 Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media.
5. Sathish Polu and Dr. V. Bapuji. “Analysis of DDOS Attack Detection in Cloud Computing Using Machine Learning Algorithm”, Tuijin Jishu/Journal of Propulsion Technology, Vol. 44, No.5, Pages:2410-2418, ISSN:1001-4055, December2023. https://www.propulsiontechjournal.com/index.php/journal/article/view/2978/2042

6. Fortuna, P., & Nunes, S. 2019 A Survey on Automatic Detection of Hate Speech in Text.
7. Zhang, Z., Robinson, D., & Tepper, J. 2020 Detecting Hate Speech on Social Media: An Analysis of Multilingual Datasets and Models.
8. D. S. Reddy, V. Bapuji, A. Govardhan and S. S. V. N. Sarma, "Sybil attack detection technique using session key certificate in vehicular ad hoc networks," *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, Chennai, 2017, pp. 1-5, https://ieeexplore.ieee.org/abstract/document/8186733/
9. Sathish Polu and Dr. V. Bapuji, "Mitigating DDoS Attacks in Cloud Computing Using Machine Learning Algorithms", The Brazilian Journal of Development ISSN 2525-8761, published by Brazilian Journals and Publishing LTDA. (CNPJ 32.432.868/0001-57) Vol.No.10, Pages:340-354, January 2024. https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/66109
10. MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. 2021 Hate Speech Detection: Challenges and Solutions.
11. Mishra, P., Yannakoudakis, H., & Shutova, E. 2022 Tackling Online Abuse: A Survey of Automatic Abuse Detection Methods.
12. Wei, J., Qian, H., & Song, Y. 2023 Towards Robust Detection of Hate Speech with Multi-Stage Machine Learning.