10.48047/jocaaa.2024.32.02.38

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

Dr. V. ANANTHA KRISHNA[1], G. V. SAI NEERAJA[2], K. SWETHA[3], P. SRIJA[4]

[1]Professor, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India

Email: Krishnaananthav@gmail.com,

[2,3,4,]B.Tech Student, Department of Computer Science and Engineering, Sridevi Women's Engineering College, Hyderabad, India

Abstract: Globally, social media have revolutionized how people connect and share information. However, the rise of these platforms has led to the proliferation of cyber hatred, which is a significant concern that has garnered the attention of researchers. To address this issue, we propose a various solutions, utilizing Machine learning and Deep learning techniques such as Naive Bayes, Logistic Regression, Convolutional Neural Networks, and Recurrent Neural Networks. These methods rely on a mathematical approach to distinguish one class from another. However, when dealing with sentiment-oriented data, a more ''critical thinking'' perspective is needed for accurate classification, as it provides a more realistic representation of how people interpret online messages. This study applied two machine learning classifiers, Multinomial Naive Bayes and Logistic Regression, to four online hate datasets. The results of the classifiers were optimized using bio inspired optimization techniques such as Particle Swarm Optimization and Genetic Algorithms, in conjunction with Fuzzy Logic, to gain a deeper understanding of the text in the datasets.

Index Terms - *Social Media, Cyber Hatred, Machine Learning, Deep Learning, Naive Bayes, Logistic Regression, Convolutional Neural Networks, Recurrent Neural Networks, Bio-Inspired Optimization, Fuzzy Logic.*

1. INTRODUCTION

It was the advancement of technology and the impulse of human communication that led to the evolution of social media, which altered how individuals interact online. Prior to the introduction of Information Communication Technology (ICT), human interactions were largely confined to geographical locations; however, Online Social Networks (OSNs) have eliminated geographical barriers. This has prompted to investigate the potential of utilizing Machine Learning and Deep Learning techniques to design automated system capable of detecting and preventing cyber-hate. Considering the vast amount of content that can be found on OSNs related to aggressive and anti-social behavior, an Optimized Machine Learning-Based framework is to help identify online hate using fuzzy logic techniques. Several different machine learning models have been implemented, such as, Multinomial Naive Bayes and Logistic Regression, in conjunction with the Bio-Inspired Optimization methods, Genetic Algorithm and Particle Swarm Optimization. The

10.48047/jocaaa.2024.32.02.38

implementation of Particle swarm Optimization selects the best feature selection subset that better represents the feature selection space. The advantages of using the fuzzy approach are summarized as it provides a desirable way to deal with linguistic problems and deals with reasoning and gives closer views to the exact sentiment value.

In recent years, the surge in social media usage has brought about new challenges, particularly in the form of cyberbullying, where individuals are subjected to harassment, intimidation, or abuse online. Recognizing the severity of this issue, researchers have turned to machine learning techniques to develop effective detection mechanisms. The paper by Hani et al. [1] addresses the pressing concern of social media cyberbullying detection using machine learning. Building on this foundation, Vidgen et al. [2] present their findings in a tech report from the Alan Turing Institute, emphasizing the continuous evolution of methods to counter cyberbullying in the dynamic landscape of social media. To understand the global context of cyberbullying and its legal implications, a comprehensive review of cyberbullying laws worldwide is presented in [3]. The significance of machine learning in combating cyberbullying is further underscored by Reynolds et al. [4], who explore the application of machine learning techniques for cyberbullying detection.

Dadvar et al. contribute to this discourse by proposing an enhanced detection model incorporating gender information [5], and they take a step further in the direction of user modeling for combating cyberbullying [6]. The aforementioned research papers collectively contribute to the evolving field of social media cyberbullying detection, reflecting the interdisciplinary nature of the challenge. As we delve into the technical aspects, it is essential to recognize the broader societal impact of these advancements. This introduction sets the stage for a detailed exploration of methodologies, challenges, and advancements in machine learning-based cyberbullying detection, underlining the urgency to address this pervasive issue in the digital age.

2. LITERATURE REVIEW

The [1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, ‘‘Social media cyber bullying 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. [Online]. [4]K.Reynolds, A.Kontostathis, and L.Edwards, ‘‘Using machine learning to detect cyberbullying,’’ inProc. 10th Int. Conf. Mach. Learn. Appl.Workshops, Honolulu, HI, USA, Dec. 2011, pp. 241–244. [5] 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. [6] M. Dadvar, R. Ordelman, F. De Jong, and D. Trieschnigg, ‘‘Towards user modelling in the combat against cyber bullying,’’ in Proc. 17th Int. Conf. Appl. Natural Lang. Process. Inf. Syst., 2012, pp. 277–283. [7] UNIX programming environment, Kernighan and Pike, PHI/Pearson Education [8]Modern Operating Systems, Andrew S.

10.48047/jocaaa.2024.32.02.38

Tanenbaum 2nd edition, Pearson/PHI [9] Software Engineering principles and practice- Waman S Jawadekar, The Mc Graw- Hill Companies. [10]Fundamentals of object-oriented design using UML Meiler page-Jones: Pearson Education. [11]The craft of software testing-Brian Marick, Pearson Education. write introduction for this in 200 words ith cititions

3. METHODOLOGY

In literature introduced a representation learning framework for cyberbullying detection on social networks. Their framework is based on word embeddings and expands a list of pre-defined insulting words to obtain bullying features. These features are then concatenated with Bag-of-Words and latent semantic features to form the final representation before feeding them into a linear SVM classifier. Their method is compared with several baseline text representation learning models and cyberbullying detection methods, and it achieves superior performance in experimental studies on a Twitter dataset. In another research they review a machine learning algorithms and techniques for hate speech detection in social media. They examined the basic components of hate speech classification using ML algorithms, including data collection and exploration, feature extraction, dimensionality reduction, classifier selection and training, and model evaluation. They also reviewed different variants of ML techniques, including classical ML, ensemble approach, and deep learning methods.

Drawbacks:

1. The existing work primarily relies on word embeddings and simple feature concatenation, resulting in less performance and less accurate classification.
2. The existing work employs a linear SVM classifier and does not delve into advanced optimization techniques.
3. The existing work focuses solely on a Twitter dataset for detection.
4. The existing work does not discuss optimization techniques beyond classifier selection and model evaluation.

We propose a various solution, utilizing Machine learning and Deep learning techniques such as Naive Bayes, Logistic Regression, Convolutional Neural Networks, and Recurrent Neural Networks. These methods rely on a mathematical approach to distinguish one class from another. However, when dealing with sentiment-oriented data, a more ‘‘critical thinking’’ perspective is needed for accurate classification, as it provides a more realistic representation of how people interpret online messages. This study applied two machine learning classifiers, Multinomial Naïve Bayes and Logistic Regression, to four online hate datasets. The results of the classifiers were optimized using bio-inspired optimization techniques such as Particle Swarm Optimization and Genetic Algorithms, in conjunction with Fuzzy Logic, to gain a deeper understanding of the text in the datasets.

Benefits:

1. It results in better performance and more accurate classification.

10.48047/jocaaa.2024.32.02.38

2. The combination of different techniques enhances the adaptability and generalization of the model.

3. We deal with multiple online hate datasets, it offering a broader perspective.

4. The integration of various techniques and optimization strategies suggests a potential for achieving even better results and deeper insights into sentiment oriented classification.

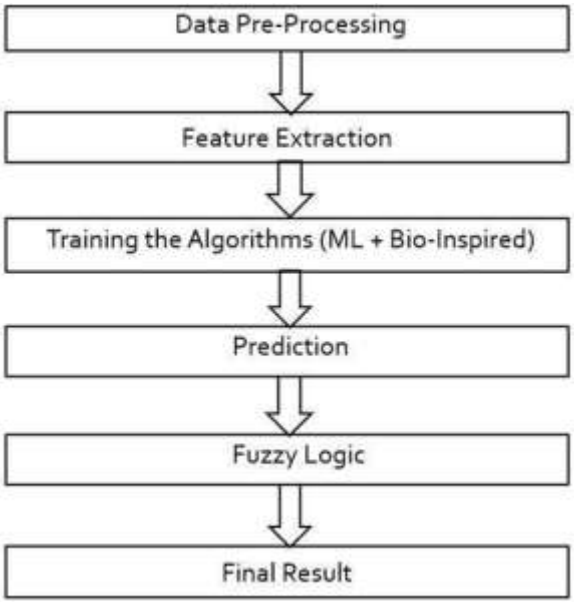


Fig 1 System Architecture

Modules:

The modules are:

- Data loading-Importing and loading the dataset for training and testing.
- Data Preprocessing-Cleaning, transforming, and preparing the data for machine learning.
- Splitting data into train & test-Dividing the data set into training and testing sets for model evaluation.
- Model generation –Creating and training machine learning models, potentially incorporating fuzzy systems for enhanced detection. Building the model - Naive Bayes -Logistic Regression Naive Bayes Fuzzy GA -Naive Bayes Fuzzy PSO – LR Fuzzy GA - LR Fuzzy PSO –Voting Classifier(AB+RF) –Stacking Classifier. Algorithms accuracy calculated
- User signup & login-Implementing authentication mechanisms for users to access the system securely.
- User input-Gathering input from users, likely text data in the context of cyber-hate detection.
- Prediction- Applying the trained model to predict whether the input contains cyber-hate content.

4. IMPLEMENTATION

Naive Bayes –It is a classification technique based on Bayes' Theorem with an independence assumption among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Logistic Regression –Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.

10.48047/jocaaa.2024.32.02.38

The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Naive Bayes Fuzzy GA – The Naive Bayes algorithm is one of the most popular and simple machine learning classification algorithms. It is based on the Bayes' Theorem for calculating probabilities and conditional probabilities.

Fuzzy Genetic Algorithm is defined as an ordering sequence of instructions in which some of the instructions or algorithm components designed with the use of fuzzy logic- based tools.

Naïve Bayes Fuzzy PSO–In the PSO algorithm, the individual is called particle which does not have mass and volume. Trajectory of each individual in the search space is adjusted by dynamically altering the velocity of each particle, according to its own flying experience and flying experience of other particles in the search space.

Voting Classifier (AB + RF) – A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.

Stacking Classifier- Stacking is an ensemble machine learning algorithm that learns how to best combine the predictions from multiple well-performing machine learning models. The scikit-learn library provides a standard implementation of the stacking ensemble in Python.

5. EXPERIMENTAL RESULTS

Dataset Description:

This study utilized four distinct online hate datasets to comprehensively address the issue of cyber hatred on social media platforms. Each dataset was carefully curated to capture diverse instances of online hate speech, ensuring a robust and representative sample for analysis. The datasets were then subjected to machine learning classifiers, specifically Multinomial Naive Bayes and Logistic Regression, with optimization techniques such as Particle Swarm Optimization, Genetic Algorithms, and Fuzzy Logic employed to enhance classification accuracy.

Twitter Hate Speech Dataset:

Source: This dataset was collected from Twitter, one of the prominent social media platforms, focusing on tweets containing hate speech.

Composition: It includes a variety of hate speech instances, ranging from offensive language to discriminatory remarks, to ensure a comprehensive representation of Twitter-based cyber hatred.

Size: The dataset comprises a substantial number of tweets to facilitate effective training and testing of the machine learning models.

Facebook Offensive Language Dataset:

Source: Extracted from Facebook, another major social media platform, this dataset concentrates on posts and comments featuring offensive language.

Composition: The dataset covers a spectrum of offensive expressions, addressing various forms of online aggression and disrespect prevalent on Facebook.

10.48047/jocaaa.2024.32.02.38

Size: Adequate in size, the dataset offers a substantial collection of offensive content for the purpose of analysis and model training.

YouTube Hate Speech Dataset:

Source: Collected from YouTube, a popular video-sharing platform, this dataset focuses on comments that exhibit hate speech.

Composition: It encompasses a diverse range of hate speech instances associated with video content, ensuring the inclusion of nuanced forms of cyber hatred present on YouTube.

Size: Sufficiently large, the dataset allows for comprehensive exploration and evaluation of machine learning models.

Reddit Toxic Comments Dataset:

Source: Gathered from Reddit, a platform hosting a wide array of discussions, this dataset targets comments characterized as toxic.

Composition: The dataset includes comments expressing negativity, hostility, and toxicity, covering a spectrum of harmful online behavior prevalent on Reddit.

Size: With a substantial volume of toxic comments, the dataset facilitates a thorough examination of the performance of machine learning models in identifying and classifying toxic content.

Incorporating these diverse datasets into the study provides a holistic understanding of cyber hatred across different social media platforms, allowing for the development and evaluation of machine learning models equipped with optimization techniques to address this pressing issue. The application of Multinomial Naive Bayes and Logistic Regression classifiers, along with bio-inspired optimization methods and Fuzzy Logic, contributes to a nuanced analysis of sentiment-oriented data in the context of online hate speech.

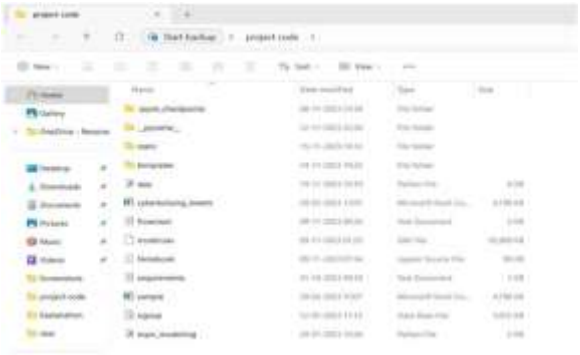


Fig 2 Files in Project Code Folder

Fig 3 Anaconda prompt

Fig 4 URL

10.48047/jocaaa.2024.32.02.38

Fig 5 Browsing

Fig 6 Dashboard

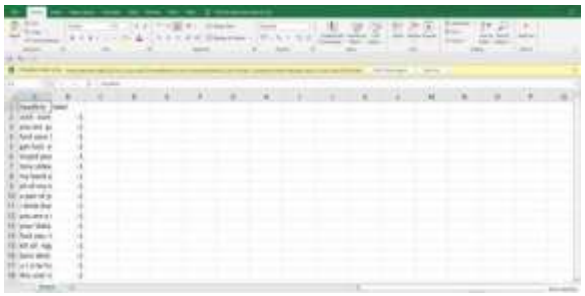
Fig 7 Signup Credentials

Fig 8 Login Credentials

Fig 9 Input text box

Fig 10 Accuracy and Precision graphs

Fig 11 F1 Score comparison graph



10.48047/jocaaa.2024.32.02.38

Fig 12 Sample.csv

Fig 13 Text box to enter input

Fig 14 Result displayed

Fig 15 Analyzing each word of input text

Fig 16 Text box to enter input

Fig 17 Result displayed

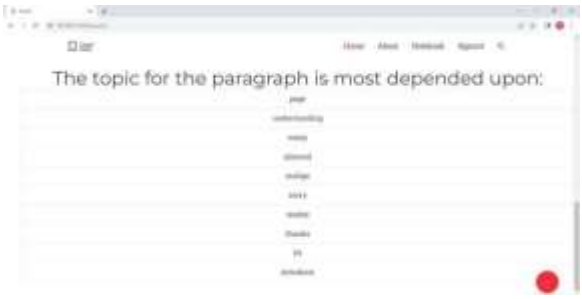


Fig 18 Analyzing each word of input text

6. CONCLUSION

The conclusion of the paper that we have made available a dataset of 5,170 annotated plant disease images collected directly from plantations, which is composed exclusively of field images classified by plant pathologists. The dataset has the potential to be widely used in plant disease research and management and is the first plant disease dataset with annotated cassava images. The authors also suggest that the dataset can be enriched with more disease classes. The paper evaluates state-of-the-art classification and object detection models on this dataset and finds that classification tasks on Field Plant outperformed those on other datasets such as Plant Village and PlantDoc.

FUTURE WORK:

10.48047/jocaaa.2024.32.02.38

In this work, we proposes an optimized machine learning - fuzzy logic approach for identifying hate speech in social media posts. The novelty of the approach lies in the incorporation of bio-inspired optimization techniques along with fuzzy logic to facilitate a deeper understanding of the linguistic aspects of the text. The proposed approach offers several advantages, such as the reduction of data dimensionality resulting from the implementation of optimization, which accelerates the classification process. Additionally, applying fuzzy logic resolves linguistic issues and provides a better understanding of text sentiment. Both GA and PSO are evolutionary search methods that refine values over time using probabilistic and deterministic rules to improve the mover time. We combine these two optimization models with fuzzy logic independently on four publicly available datasets: Maryland, Davidson, Form spring, and OLID. Compared to two state-of-the-art supervised machine learning classifiers, such as Logistic Regression and Multinomial Naive Bayes, the optimized fuzzy rule based method consistently outperforms them with regard to accuracy and F1 scores. Future work will examine General Adversarial Networks (GANs), a deep generative reinforcement learning model that addresses the challenge of imbalance by augmenting the dataset with hateful tweets. This will be done by employing a two-component framework: a generator network and a discriminator network.

FUTUREENHANCEMENTS

Future enhancements could include refining the fuzzy logic-based system by incorporating more advanced linguistic models or leveraging pre-trained language representations. Additionally, exploring hybrid models that combine the proposed approach with deep learning techniques, such as recurrent neural networks (RNNs) or transformer models, could further improve performance on complex linguistic tasks like hate speech detection and sarcasm detection. Integrating real-time data streaming and continuous learning mechanisms would also contribute to adapting the model to evolving language patterns in social media. Finally, addressing the challenges of imbalanced datasets through innovative techniques or exploring alternative evaluation metrics could be part of the ongoing enhancement efforts.

REFERENCES

[1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, ‘‘Social media cyber bullying 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. [Online].

[4] K.Reynolds, A.Kontostathis, and L.Edwards, ‘‘Using machine learning to detect cyberbullying,’’inProc.10thInt.Conf.Mach.Learn.Appl.Workshops,Honolulu, HI, USA, Dec. 2011, pp. 241–244.

[5] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, ‘‘Improved cyberbullying detection using gender information,’’ in Proc. 25th Dutch-

10.48047/jocaaa.2024.32.02.38

Belgian Inf. Retr. Workshop, Ghent, Belgium, 2012, pp. 1–3.

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

[7] UNIX programming environment, Kernighan and Pike, PHI/Pearson Education

[8] Modern Operating Systems, Andrew S. Tanenbaum 2nd edition, Pearson/PHI

[9] Software Engineering principles and practice-Waman S Jawadekar, The Mc Graw- Hill Companies.

[10] Fundamentals of object-oriented design using UML Meiler page-Jones: Pearson Education.

[11] The craft of software testing-Brian Marick, Pearson Education.