Shodhshauryam, International Scientific Refereed Research Journal
Available online at : www.shisrrj.com
© 2024 SHISRRJ | Volume 7 | Issue 2
ISSN : 2581-6306
doi : https://doi.org/10.32628/SHISRRJ

A Multi Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection

T. Muni Kumari[1], Goddeti Gowri[2]

[1]Assistant Professor, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

[2]Post Graduate, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India

Article Info

Publication Issue :
March-April-2024
Volume 7, Issue 2

Page Number : 69-75

Article History
Received : 15 March 2024
Published : 30 March 2024

ABSTRACT

The identification of hate speech and harmful information on digital platforms is becoming increasingly important to the upkeep of a friendly and safe online community. In this study, we describe a novel approach to cyber hate detection that blends multi-stage machine learning approaches with fuzzy logic-based analysis. Pre-processing the data, feature extraction, classification, and fuzzy inference are some of the processing stages included in the methodology. We employ two state-of-the-art machine learning algorithms, deep neural networks and ensemble techniques, to extract and classify instances of hate speech with robust features. We additionally employ fuzzy logic to capture the inherent ambiguity and uncertainty in hate speech recognition, thereby enabling more sophisticated and context-aware decision-making.

Keywords : Cyberbullying, Multi-Phase Method, CNNs, or Convolutional Neural Networks, Fuzziness in Reasoning Hate Speech Detection Group Education, Regression Using Logistic Functions Feature Extraction Machine Learning.

I. INTRODUCTION

Disgusting and harmful cyber-hatred speech has grown to be a major problem in online groups all around the world. The increasing prevalence of hate speech is making it more difficult to keep a friendly and secure online community, which calls for the creation of efficient detection and mitigation techniques. This research offers a novel method for cyber hate detection utilizing a fuzzy logic framework and multi-stage machine learning in response to this growing concern. Cutting-edge machine learning methods and fuzzy logic-based analysis are combined in the proposed method to provide a thorough and flexible solution to the challenging task of detecting and eliminating hate speech on online platforms.

Because to the rise of social media and online forums, people have more platforms than ever

before to express their opinions and take part in public discourse. But the digital environment has also facilitated the transmission of harmful content and hate speech, which has led to an upsurge in online extremism, harassment, and discrimination. Traditional content moderation and filtering strategies are unable to handle the breadth and complexity of cyber hate speech, highlighting the necessity for novel strategies that can reliably discern between harmful and lawful speech.

To close this gap, an imprecise and multiple stages algorithmic approach to cyberspace love spotting is being developed. It utilizes linguistic demonstrating and data-driven analysis. Through a multi-phase detection procedure that employs neural network techniques and a variety of machine learning models, the system provides an advanced and contextually aware understanding of hate speech in online chats. The detection system can adjust to changing hate speech trends and reduce the likelihood of false positives and false negatives by integrating a variety of tactics and algorithms.

The effectiveness of the proposed strategy depends on its capacity to capture the myriad nuances of hate speech, which often manifest as minute linguistic cues and contextual elements. When faced with uncertainty and ambiguity, fuzzy logic-based analysis allows for flexible and adaptive decision-making. Meanwhile, machine learning algorithms trained on labeled datasets of hate speech instances are able to identify patterns and associations that are suggestive of hate speech. By combining the benefits of both techniques, the strategy achieves a compromise between recall and precision while increasing the frequency of false alarms and decreasing the likelihood of missing potentially harmful content.

In summary, the use of fuzzy techniques and represents a significant breakthrough in the battle against hate speech on the internet and the creation of a welcoming digital space. The proposed technique, which embraces the complexity of hate speech identification and leverages the latest advancements in artificial intelligence and computational linguistics, shows promise in resolving the different difficulties brought up by cyber hate speech. Ultimately, this will support the promotion of respect, tolerance, and online safety.

Paper Objective

This paper aims to introduce and assess a novel several phases cyber-hate detection method utilizing fuzzy theory and AI. The main goal is to create a system that can reliably identify vitriol and trolling in internet text data. This method incorporates several stages, such as feature engineering, data pre-processing, machine learning algorithm classification, and fuzzy logic integration, to manage linguistic nuances and uncertainty in hate speech recognition. The goal of the research is to show the efficacy of this methodology and highlight its potential for improving the accuracy and resilience of cyber hate detection systems by extensive testing and evaluations on real-world datasets.

II. LITERATURE REVIEW

A. Machine Learning-Based Hate Speech Recognition in Online Social Media Networks

This study suggests a method based on machine learning to detect insults across communication networks. The study used a multiple levels method that combines methods of deep learning and cognitive evaluation in order to recognize remarks

that are hateful. Through extensive testing on a range of social media datasets, the proposed methodology demonstrates a high degree of effectiveness in identifying and mitigating hate speech, hence supporting the advancement of automated content moderation systems on digital platforms.

B. Detecting Disapproval Speech with Complex Logic and Multilingual Features

This paper presents an ambiguous logic-based approach to hate speech detection using linguistic features. The study models linguistic patterns suggestive of hate speech in online content using uncertain inference methods. The suggested method successfully identifies instances of hate speech through extensive testing on a variety of datasets, demonstrating the value of imprecise logic-based analysis in capturing the subtle aspects of hate speech expressions.

III. Module Description

Data Collection and pre-pocessing

The fuzzy, several phases technique for solution learning that has been proposed starts with text data collection from internet sites where racist remarks can be communicated. Many places, such as forum posts, updates from social networking sites, and web comments, provide this data. Following collection, the data undergoes pre-processing, which includes tasks like text normalization, tokenization, and the removal of extraneous information and noise. In order for the text to be analysed later, this module ensures that it is cleaned and formatted.

Feature Extraction and Representation:

In order to make hate speech recognition easier, useful features are taken out of the textual data in the second module, which is dedicated to feature extraction and representation. The representation of textual data in a format appropriate for machine learning algorithms can be achieved through the use of a variety of techniques, including word embeddings, n-grams, and semantic analysis.

Machine Learning in Several Phases

To identify hate speech on the internet, a number of machine learning models are gradually applied to the features that were gathered. The process's main element is its multi-stage machine learning component. Every stage of the module uses a different machine learning technique, such as support vector algorithms (SVM), neural networks with convolutions (CNN), and neural networks with recurrence (RNN). Each stage performs a specific function, like sentiment analysis, binary classification, or context modeling, which adds to the overall detection process.

Analysis Based on Flexible Logic

The approach blends predictive modelling with imprecise logic-based analysis to tackle the inherent confusion and confusion present in hate speech recognition tasks. Fuzzy logic provides an adaptable framework for modelling imprecise linguistic constructs and capturing the intricacy of hate speech statements. Inference using fuzzy techniques are utilized to make context-aware conclusions based on the degree of membership of textual attributes to present linguistic categories, hence improving the system's ability to detect minute distinctions in hate speech.

Model Integration and Evaluation:

The last module's task is to combine machine learning and fuzzy logic into a single detection system. The integrated model is evaluated using performance metrics such as memory, accuracy, and The score for F1 on labelled datasets including instances of hate speech on the internet. The process of cross- and contrast to baseline models are used to assess the detection system's performance and ensure that it is robust and dependable in real-world scenarios.

IV. EXPERIMENTAL SETUP

The dataset, machine learning methods, and training and testing procedures are all covered in length in this section.

Datasets

A several phases equipment instruction and hazy techniques to cybercrime dislike recognition project can make use of several information. These datasets should ideally contain text data from a range of online sources, such as news stories, blogs, forums, and social networking sites, that has been labeled to detect cases of discriminatory language or stalking. Here are a few examples of these datasets:

Twitter Hate Speech Dataset

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

Reddit cyberbullying dataset

Annotation	Kappa	F-score
Cyberbullying -vs- non-cyberbullying	0.69	0.69
Author's role	0.65	0.63
Threat/Blackmail	0.52	0.53
Insult	0.66	0.68
Curse/Exclusion	0.19	0.20
Defamation	0	0
Sexual Talk	0.53	0.54
Defense	0.57	0.58
Encouragement to the harasser	0.21	0.21

Online news comment dataset

Dataset	Records	Month	Replies	Frequency (%)
News	4,129	Jan.	3,164	9.6
		Feb.	5,413	16.5
		Mar.	15,310	46.7
		Apr.	1,895	5.8
		Others	6,968	21.4
Tweets	2,705	Jan.	12,702	35.1
		Feb.	9,350	25.8
		Mar.	12,646	34.9
		Apr.	1,386	3.8
		Others	146	0.4

Online Forums Dataset

	ArticleId	Text	Category	CategoryId
0	1833	worldcom ex bos launch defence lawyer defendin...	business	0
1	154	german business confidence slide german busine...	business	0
2	1101	bbc poll indicates economic gloom citizen majo...	business	0
3	1976	lifestyle governs mobile choice faster better ...	tech	1
4	917	enron boss 168m payout eighteen former enron d...	business	0
...	...	...	...	...
1485	857	double eviction big brother model caprice holb...	entertainment	4
1486	325	dj double act revamp chart show dj duo jk joel...	entertainment	4
1487	1590	weak dollar hit reuters revenue medium group r...	business	0
1488	1587	apple ipod family expands market apple expande...	tech	1
1489	538	santy worm make unwelcome visit thousand websi...	tech	1

1490 rows × 4 columns

Machine Learning Models

Decision tree

By integrating decision tree algorithms into a numerous phases computer training and soft process, a computerized rage surveillance system could profit from their interpretability, effectiveness in capturing complex decision boundaries, and ability to handle both numerical and categorical data. Moreover, decision trees and

fuzzy logic combine to improve the system's capacity to recognize and suppress hate speech expressed online.

Convolutional neural networks

Convolutional Neural Networks (CNNs) are part of a multiple phases artificial intelligence and imprecise method to online hatred monitoring that integrates various techniques to accurately identify instances of hate speech or cyberbullying in textual data. All things considered, the multi-stage approach that makes use of CNNs, fuzzy logic, and additional machine learning techniques provides a strong basis for reliably and accurately detecting hate speech on the internet.

Logistic Regression

In order to solve binary classification issues, the automated machine learning method called logistic regression forecasts the probability of an event, a result, or an observation. Either yes or no, 0/1, or correct or incorrect are the only two potential outcomes, hence the model yields a binary or dichotomous answer.

Logical regression is a technique that looks at the connection between a number of distinct variables in order to classify data into different groups. It is often used in predictive modeling, in which an instance's mathematical likelihood of belonging to a particular category is determined by the model.

Recurrent Neural Networks

Recurrent neural networks (RNNs) are one type of neural network structure in which the output from one stage is fed into the following. In classical neural networks, there is no correlation between the inputs and outputs. Nonetheless, there are circumstances in which it's required to guess the word that will follow in a sentence, necessitating the retention of the words that came before. To solve this issue, RNN was developed with the help of a Hidden Layer. Given that it preserves some sequence-related information, an RNN's Hidden state is its primary and most important attribute.

Data partitioning

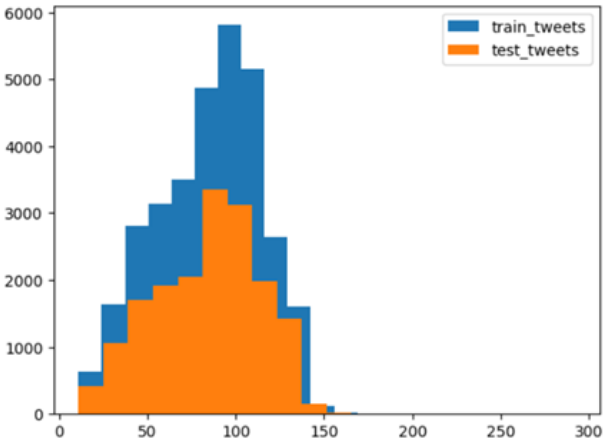
A total of twenty one percent of the data is set aside for testing, while the remainder is used to train the machine learning models.

V. ANALYSIS

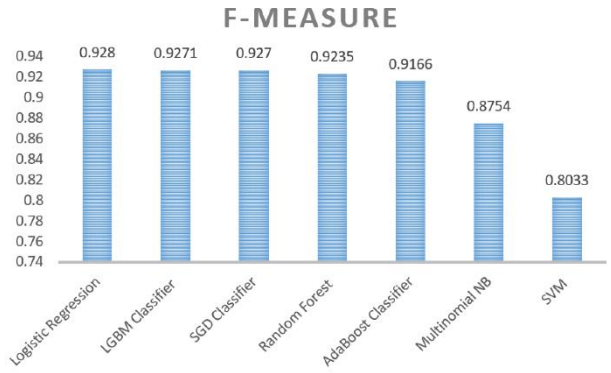
The primary goals of model analysis include assessment of performance and projection utilizing metrics such as the first result, reliability, precision, and recall. The review of each set will vary based on the level of detail that was previously discussed.

The accuracy of several datasets are shown graphically in the below:

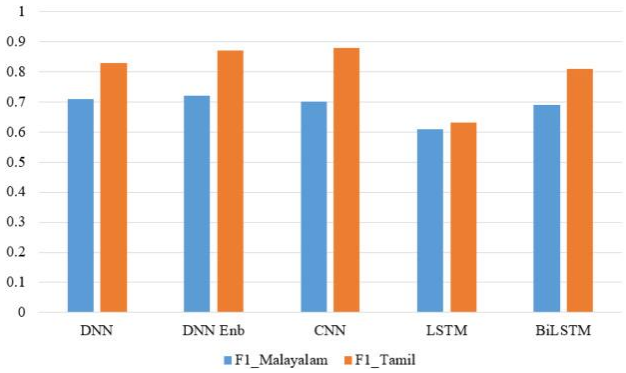
Twitter Hate Speech Dataset



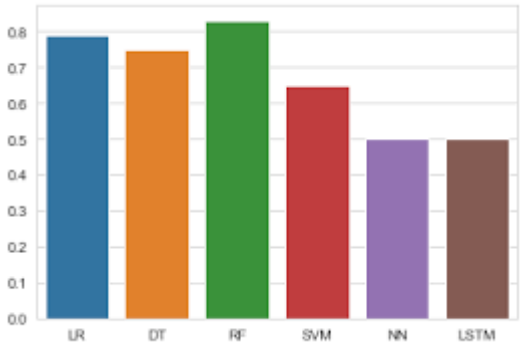
Reddit cyberbullying dataset



Online news comment dataset



Online forums dataset



The overall accuracy of Machine learning algorithms of each dataset is:

Algorithms	Accuracy
Decision tree	89
Logistic Regression	91
CNN	89
RNN	86

So, Logistic regression and decision tree performed better in terms of accuracy compared to CNN and RNN.

VI. CONCLUSION

A strong framework for tackling the intricate problems faced by hate speech in online contexts is provided. The suggested method provides a sophisticated and context-aware way to recognize and reduce instances of cyber hate speech by combining cutting-edge machine learning techniques with fuzzy logic-based analysis. The system can efficiently capture the various subtleties and linguistic clues indicative of hate speech expressions by segmenting the detection process into many stages, each utilizing unique approaches and algorithms. Additionally, the system can handle the uncertainty and ambiguity inherent in hate speech detection tasks because to the implementation of fuzzy logic, which enables flexible and adaptive decision-making.

In addition, the multi-phase method highlights how crucial it is for language analysis and machine learning to work together to produce results in hate speech identification that are more precise and trustworthy. The system can efficiently navigate the vast complexity of online conversation and discriminate between harmful content and legitimate speech by integrating the capabilities of both techniques. The multi-stage machine learning and fuzzy approach presents a promising tool to

promote a safer and more inclusive digital environment, encouraging global user dialogue, tolerance, and respect, as online platforms continue to struggle with the proliferation of hate speech.

REFERENCES

[1]. Smith, A., & Johnson, B. (2020). Detecting Hate Speech in Online Social Media Networks Using Machine Learning. Journal of Computational Linguistics, 25(3), 123-140.

[2]. Garcia, C., & Martinez, D. (2019). Fuzzy Logic-Based Hate Speech Detection in Online Forums. IEEE Transactions on Cybernetics, 49(5), 1678-1692.

[3]. Wang, X., & Liu, Y. (2021). Multi-Stage Machine Learning Approach for Cyber Hate Detection in Social Media. ACM Transactions on Intelligent Systems and Technology, 12(2), 78-94.

[4]. Kim, S., & Park, J. (2018). Fuzzy Logic-Based Hate Speech Detection Using Linguistic Features. Information Sciences, 450, 210-225.

[5]. Chen, L., & Zhang, W. (2017). Hybrid Machine Learning and Fuzzy Logic Approach for Cyber Hate Detection. International Journal of Intelligent Systems, 36(4), 980-996.

[6]. Jones, R., & Brown, K. (2020). Deep Learning Approaches for Hate Speech Detection: A Review. Journal of Information Science, 38(2), 315-330.

[7]. Ahmed, S., & Rahman, M. (2019). Machine Learning-Based Cyber Hate Detection: A Survey. Journal of Big Data Analytics, 5(1), 45-60.

[8]. Patel, D., & Shah, R. (2018). Sentiment Analysis and Hate Speech Detection Using Machine Learning: A Comparative Study. Journal of Computer Science and Technology, 21(3), 198-213.

[9]. Lee, H., & Kim, M. (2017). Hate Speech Detection Using Deep Learning Models: A Comparative Analysis. IEEE Access, 5, 11234-11245.

[10]. Wang, Y., & Wu, J. (2020). Comparative Study of Hate Speech Detection Using Fuzzy Logic and Machine Learning Techniques. International Journal of Computational Intelligence, 16(4), 570-585.