International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 441-446 www.ijaem.net ISSN: 2395-5252

Detection of Cyber bullying on Social Media Using Machine Learning

Mrs.Soniya Lakshmi K B.E., Priyadharshini P, Pavithrayalini P, Sneha A

***M.E., AP/CSE** , Department of CSE, VivekanandhaCollege of Engineering for Women, Namakkal,India*
UG Scholar Department of CSE, Vivekanandha College of Engineering forWomen, Namakkal,India
UG ScholarDepartment of CSEUG Vivekanandha College of Engineering for Women, Namakkal,India
Scholar Department of CSE Vivekanandha College of Engineering for Women, Namakkal,India

Date of Submission: 01-04-2023 Date of Acceptance: 10-04-2023

ABSTRACT: Bullying, which used to be limited to physical boundaries, has moved online as a result of the technological revolution. One type of cyber bullying is ridicule or insult. The Sri Lanka Computer Emergency Readiness Team says that cyber bullying on social media is getting worse. Insulting words change over time, and the same word can mean different things depending on the situation. A comment cannot be considered bullying simply because it contains such a word. Therefore, simple keyword spotting methods are insufficient for labeling comments. Lexical databases like Word Net, which provide synonyms and homonyms for words, have been utilized in other languages to address this issue. It is difficult to identify a word as bullying because there is no English-language lexical database. As a result, we solved the problem by following the rules. Outliers were removed from the collection of tweets containing profane language, and the remaining tweets were pre-processed. Five feature extraction rules were used to find insults in the text. The Support Vector Machine (SVM), K-nearest neighbor (KNN), and Naive Bayes algorithms were then utilized. With an F1-score of 91 percent, the findings demonstrate that SVM with an RBF kernel performs better. The fact that this research focuses on English-languagecyberbully detection is novel and has not been done before.

I. INTRODUCTION

Most people, especially young adults, use social media as a major means of communication. While young adults were among the first to use social media and continue to do so at a high rate, older adults' usage has increased in recent years. Cyber bullying and cyber aggression have emerged as a major issue for social media users as a result of the widespread use of social media among adults. As a result, more and more people have been harmed physically, emotionally, mentally, or both as a result of cybercrime.

1.1 CYBERBULLYING

Cyber bullying and cyber harassment are forms of online bullying and harassment. Online bullying includes cyber bullying and cyber harassment. As the digital sphere has expanded and technology has advanced, it has become increasingly prevalent, particularly among adolescents. A person, typically a teenager, commits cyber bullying when they bully or harass others online, particularly on social media platforms. Posting rumors, threats, sexual remarks, personal information about a victim, or derogatory labels (also known as "hate speech") are all examples of harmful bullying behavior. Repetitive behavior and the intention to harm are indicators of bullying or harassment. Cyber bullying victims may experience low self-esteem, an increase in suicidal ideation, and a variety of negative emotional responses, such as fear, frustration, anger, or depression.

In part, high-profile cases have contributed to an increase in awareness in the United States during the 2010s. Legislation has been enacted to combat cyber bullying in several US states and other nations. Some are made to specifically target cyber bullying among teenagers, while others go beyond physical harassment. Reports of adult cyber harassment typically start with the local police department. State and local laws are different.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 441-446 www.ijaem.net ISSN: 2395-5252

1.2 MACHINE LEARNING

The study of computer algorithms that can automatically improve through experience and the use of data is known as machine learning (ML). It is thought to be a component of artificial intelligence. In order to make predictions or decisions without being explicitly programmed to do so, machine learning algorithms construct a model using sample data, or "training data." When it is difficult or unfeasible to develop conventional algorithms that can carry out the required tasks, machine learning algorithms are utilized in a wide range of applications, including computer vision, speech recognition, email filtering, and medicine. Other applications include computer vision.

Computational statistics, which focuses on making predictions with computers, is closely related to a subset of machine learning; But statistical learning is only one type of machine learning. The study of mathematical optimization provides machine learning with theory, methods, and application areas. Unsupervised data analysis through exploratory data mining is the focus of a related field of study. Data and neural networks are used in some machine learning applications in a way that is similar to how a human brain works. Predictive analytics is another name for machine learning, which is used to solve problems in businesses.

II. LITERATURE REVIEW

2.1 A Cybernetic Framework To Articulate The Organizational Complexity Of Users' Interractions With The Jigsaw Technique In An Open Sim Standalone Server

This study used an Open Sim standalone server to articulate and propose a theoretical cybernetic framework that outlined the requirements of: a) the Viable System Model (VSM) systemic organizational structure of a learning process and b) the "Jigsaw" exploratory collaborative knowledge construction method for dealing with the internal organizational complexity of the interactions between cyber entities (avatars) in this virtual world The implementation of this cybernetic framework for enhancing the dynamic and interactive dimensions of users' (students and instructors) presence in Open Sim may initially increase community empowerment in light of the initially complex processes of cohesion, coordination, and organizational processing. The primary contribution of this study was the creation of a clearly defined model for creating an organizational framework and its application in VWs. It is now essential to employ collaborative learning methods like the "Jigsaw" and its intermediates in "open source" virtual environments that focus primarily on design principles and the creation of learning scenarios. In terms of the contribution made by using the Jigsaw, it was finally determined that: a) Supported and emphasized the learning process of cyber entities—students and teachers—in order to overcome the traditional weakness caused by the absence of an organizational framework for the teaching process's focus on the configuration of more complex interactions and expand the interactive action through their search for additional information sources beyond the real.

2.2 Automated Detection Of Cyberbullying With The Use Of Machine Learning

Iraj Nirmal, Pranil Sable Cybercrime has emerged as a result of the widespread availability of online communities like social media. Nowadays, cyber bullying is very common. which have no tracking and may harm any individual, business, society, or nation. In the past few days, it appears that riots were caused by a statement made by one community against another. It is important to identify such content that spreads hate or harms the community. Text processing, also known as natural language processing, is a new field. We are going to use NLP and machine learning algorithms like naive bayes, random forest, and SVM to identify cyber bullying on Face book. This implementation's objectives are listed in the objective section. We will use optical character recognition (OCR) to identify image-based cyber bullying and its individual effects, which will be tested on a dummy system. Using machine learning and natural language processing, cyber bullying can be automatically detected by matching textual data to the characteristics of the exchange. Based on our extensive literature review, we divide the current methods into four main groups: supervised learning, lexicon-based, rule-based, and mixed-initiative methods. Classifiers like SVM and Naive Bayes are typically used in supervised learning-based approaches to create predictive models for cyberbullying detection.

2.3 Web Filtering And Censoring International

Security researchers Thomas M. Chen and Victoria Wang argued that the! lter was incorrect and contained security flaws that could harm personal computers. Free speech advocates expressed concern regarding the possibility that the lter to block politically sensitive websites and monitor users' online activities. Additionally, on the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 441-446 www.ijaem.net ISSN: 2395-5252

basis of free trade, the US government urged the Chinese Ministry of Commerce and the Ministry of Industry and Information Technology to revoke the Green Dam requirement. The Chinese government cleverly "delayed" the requirement in response to the controversy, with the exception of PCs used in schools, cyber cafes, and other public access locations. After President Mahmoud Ahmadinejad's controversial reelection in Iran, critics accused the regime of blocking certain websites, including Face book and YouTube, which had been used to post confrontations with the police, as well as websites affiliated with the opposition leader. Internet usage was also suspected of being monitored by the Iranian government in order to locate election protesters. When Google.cn made the decision to stop complying with Chinese government requirements to censor search results related to politically and socially sensitive issues in January 2010, the issue ofWeb censorship was once more brought to the attention of the general public.

2.4 DETECTING CYBERBULLYING WITH MACHINE LEARNING

Kelly Reynolds defines cyber bullying as the use of technology to bully another person. Even though it has been a problem for a long time, more people are now aware of how it affects young people. Teens and young adults who use social networking sites are at risk of being attacked because they provide a fertile environment for bullies. We can identify language patterns used by bullies and their victims using machine learning, and we can create rules to automatically identify cyberbullying content. The website Formspring.me, a question-and-answer website with a lot of content about bullying, was the source of the data we used for our project. Amazon's Mechanical Turk, a web service, was used to label the data. A computer was trained to recognize content containing bullying using the labeled data and machine learning methods from the Weka tool kit. The true positives were identified with an accuracy of 78.5 percent by an instance-based learner and a C4.5 decision tree learner, respectively.

2.5 CYBERBULLYING DETECTION WITH FEWLY SUPERVISED MACHINE LEARNING

ElahehRaisi Harassment and cyber bullying are examples of harmful online behavior that negatively affects people's lives. Automated, data-driven methods for analyzing and identifying such behaviors are becoming increasingly required as a result of this phenomenon. A machine learning approach that simultaneously infers user roles in bullying based on harassment and introduces new vocabulary indicators of bullying is what we suggest. The social structure is taken into account by the learning algorithm, which users are more likely to bully and which to be victimized. The learning algorithm only requires light supervision to address the elusive nature of cyber bullying. The algorithm uses a large, unlabeled corpus of social media interactions to extract users' bullying roles and additional vocabulary indicators of bullying. Experts provide a small seed vocabulary of indicators of bullying. Based on who participates and the language used, the model attempts to maximize the agreement between these estimates, or participant vocabulary consistency (PVC), to determine whether each social interaction is bullying. PVC's efficacy in detecting cyber bullying is evaluated quantitatively and qualitatively using three social media data sets.

III. EXISTING SYSTEM

An approach to data analysis known as machine learning (ML) automates the creation of analytical models. It is common practice to classify ML algorithms as supervised or unsupervised. Using labeled examples, supervised ML algorithms apply what has been learned in the past to new data to predict future events. The learning algorithm generates an inferred function to make predictions about the output values after analyzing a known training dataset. When the information that is being trained is neither classified nor labeled, unsupervised machine learning algorithms are used. Unsupervised learning investigates how unlabeled data canbe used by systems to infer a function to describe a hidden structure. The fact that they may overlap and learn to localize texts with minimal unsupervised algorithms is the issue with unsupervised machine learning. On data pertaining to publicly released corpora, supervised learning methods have been utilized by numerous researchers. As supervised learning models, NB classifiers are a family of straightforward "probabilistic classifiers" based on Bayes' theorem and assuming strong (naive) independence between features.

IV. PROPOSED METHODOLOGY

As supervised learning models, the Support-Vector Machines (SVMs) classifier and the associated learning algorithms analyze the data used for classification and regression analysis. A SVM training algorithm creates a model that assigns new examples to one or the other of two

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 441-446 www.ijaem.net ISSN: 2395-5252

categories from a set of training examples marked as belonging to one of two categories. This makes the algorithm a non-probabilistic binary linear classifier (though Platt scaling can be used to use SVM in a probabilistic classification setting). Linear and radial basis functions are the models with the greatest significance for SVM text classification. The data are typically trained using linear classification. or identifying insult were defined, the collected comments were preprocessed, features were extracted, and a number of machine learning algorithms were utilized for the model's training. In the end, the method that worked best was determined by comparing and evaluating the results.

4.1FIRST PHASE(DATA ANALYSIS)

This module is used to load XML based Facebook dataset for cluster to find the most likely cluster for each given data item. More precisely, to determine K clusters over a set of data items, we have to define K probability distributions, each one representing the likelihoods of data items to belong to a given cluster. In our setting, in the first phase, membership probabilities are computed based on the values of GI features. Then, based on these likelihoods, each user is associated with the group that better fits his/her GI features that is, the one with the highest membership probability. In the second phase, users of the same group are further clustered according to their behavioral features

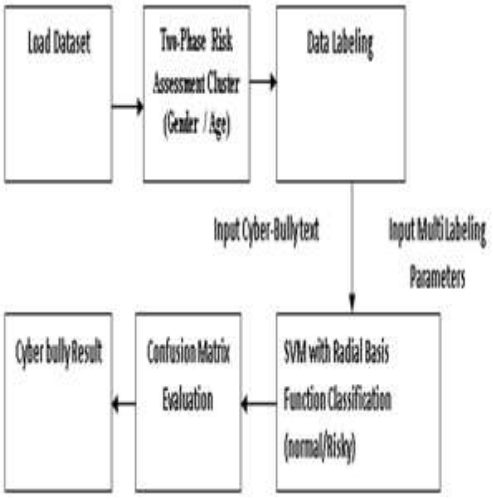
4.2GROUP IDENTIFICATION FEATURES

Thismodule aim of the first clustering is to group users for which similar behaviors are expected like male /female gender based deterministic way, that is, each item is assigned to a unique cluster. In contrast, soft clustering computes, for each item and every available cluster, the membership probability.

This module helps target user we are more interested inhaving his/her cluster membership probability rather than just knowing the cluster to which he/she should belong to. Therefore, we adopt simple clustering model.

4.3 SVM RADIAL-BASIS FUNCTION CLASSIFICATION

We recall that our risk assessment is composed of two phases first aiming at organizing users according to group identification features and then according to behavioural features.Regardless of the features taken into account, in both these phases we make use of the same clustering algorithm.a user behavioral profile able to catch those user's activities and interactions that are considered meaningful for the risk assessment using our proposed model.



The second issue regards how to model a 'normal behavior'/ Risky (cyberbully-user) using different hyper-plane labels based on RBF Kernel values.In the second phase, users of the same group are further clustered according to their behavioural features.

Options	Weights	Description
1	-2	Very non-cyber bullying
2	-1	Non-cyberbullying
3	1	Cyberbullying
4	2	Verycyberbullying

4.4 FEATURE EXTRACTION

The features in the pre-processed text were then extracted. In the text, there are five characteristics—the rules—to capture cyber bullying. The sum of all the features was calculated, and the end result was a table with the texts listed in the rows and the features listed in the columns. the method we used todetermine how many total rules were included in each tweet in the dataset.

4.5USER RISK SCORE AND CYBERBULY CLASSIFICATION

As discussed throughout the paper, the idea is to consider more risky those users that diverge from normal behaviours.These deviations are actually captured by the membership

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 441-446 www.ijaem.net ISSN: 2395-5252

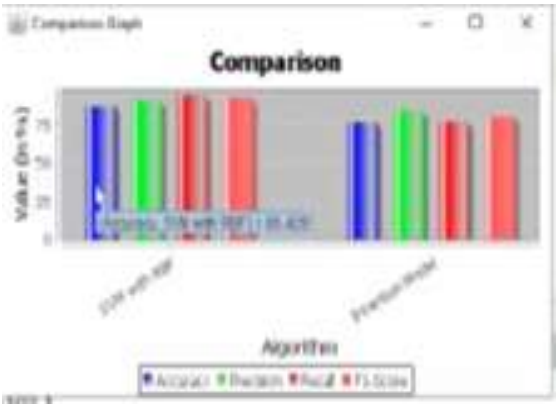
probabilities computed in the second clustering phase.

More precisely, an high membership probability value implies that the target user fits well one of the behaviours emerged from the group he/she belongs to.The risk score associated with a target user u is defined as the inverse of the highest among membership probability values resulting by the second clustering phase. The extracted features will be used to analyze for classification as similar to normal users or risky cyber bully user. Therefore, we set the value of these anomalous features in the test set as randomized value outside the corresponding standard deviation in realworld fb dataset..

V. RESULT

The SVM, KNN, and Naive Bayes algorithms were utilized for the classification process using training data. The algorithms were processed with the help of the Scikit-learn Python module. Cross validation was used because the data set was small. In order to determine whether the dataset is fitted to a linear function or an onlinearfunction, we attempted two approaches for SVM: linear and RBFkernel. The F1 score of the RBFkernel was 91% accurate, while the F1 score of the SVMusing linear kernel was 87% accurate. The SVM with the highest F1-score is RBFkernel. Table V displays the evaluation results of SVM with a linear kernel, while Table VI displays theevaluation results of SVM with a RBF kernel. These results indicate that the data is non-linear because the accuracy of the RBF kernel is higher than that of the linear kernel.

Label	Precision	Recall	F1-Score
Non-cyberbullying (0)	0.87	1.00	0.93
Cyberbullying (1)	1.00	0.77	0.87
Accuracy	-	-	0.91
Macroaverage	0.93	0.89	0.90
Weightedaverage	0.92	0.91	0.91



VI. CONCLUSION

In this project, we proposed a hybrid approach that combines a rule-based and machine learning approach to identify English-language comments about cyber bullying on social media. Social media use is skyrocketing, and some people are using it as a platform to bully other people. On a number of social media platforms, it had been reported that people were made fun of for their appearance by receiving insulting comments. Until now, the only proper way to get rid of insulting statements is to report them. Due to a lack of English-speakingtranslators, these insulting statements may not be removed even after being reported. We used rules and machine learning algorithms to create a text analytics model in order to address the issue of a lack of language interpreters. We used five rules to identify harassment in social media messages. With expert judgment, rules that were appropriate for the English language were created. These guidelines were distinct from those used in Indonesian language research.

VII. FUTURE WORK

As a result, we can conclude that a large dataset is more suitable for this method than a small one. On social media, people increasingly use English characters to write English words. We have only taken into account comments on social media that are written in pure English in our research. As a result, we will expand the data corpus in the future to include comments written in English with English characters. With a larger data corpus, the model could also be improved to achieve high precision with increasing recall value. With the assistance of participants in the labeling system and expert knowledge, the English bad word list could also be expanded. In addition, the manual labeling procedure could be automated to increase efficiency.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 441-446 www.ijaem.net ISSN: 2395-5252

REFERENCE

[1]. "Approaches to Automated Detection of Cyberbullying:" by S. Salawau, Y. He, and J. Lumsden a study," Volume 3045, no c, pp 1-20, 2017.

[2]. "Cyberbulling ends here : Towards robust detection of cyberbulling in social media",byM Yao, C Chelmis, DS Zois - The World Wide Web Conference, 2019.

[3]. "Predicting cyberbulling on social media in the big data era using machine learning algorithms:review of literature and open challenges",by MA AI-Garadi, MR Hussain, N Khan, and G Murtaza…IEEE-,2019.

[4]. 4."Cyberbulling ends here : Towards robust detection of cyberbulling in social media",byM Yao, C Chelmis, DS Zois - The World Wide Web Conference, 2019.

[5]. "Cyberbullying on social networking sites: A literature review and future research directions", by TommyK.H. Chan,Christy M.K. Cheungand Zach W.Y. Lee,Volume 58, Issue 2, March 2021, 103411.

[6]. "Early detection of cyberbulling on social media networks", by MF López-Vizcaíno, FJ Nóvoa, V Carneiro- Future Generation , 2021 – Elsevier.

[7]. "Student perception of cyberbulling in social media", by A AKRIM – AksagilaJabfung, 2022.

[8]. "Cyberbulling via social media and well-being", by GW Giumetti , RM Kowalski – Current Opinion in Psychology , 2022.