

Detecting Cyber Bullying using ML Algorithm

S Ravikumar¹, K. Udaykiran², Omkar Adabala³, Duggempudi Jagadeeswar Reddy⁴ and Adabala Siva Deepthi⁵

¹⁻⁴Department of Computer Science and Engineering, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai, Tamilnadu, India.

Email: ravikumars.086@gmail.com, kalyanapuudaykiran@gmail.com, adabalaomkar@gmail.com, jagadishd0123@gmail.com

⁵Department of Information & Communication Technology, Sastra Deemed University
Email: adabalasivadeepthi@gmail.com

Abstract—Cyberbullying is a crime in which a victim attacks a person with hatred and online threats. In order to recognize word similarity in bullies' tweets and allow use of machine learning, an ML model can be built to automatically identify social network bullying behavior. Nonetheless, several social media abuse detection approaches have been adopted, but many of them are textually focused. To recognize and avoid bullying on Twitter, a machine learning algorithm is recommended. 4 classifiers, i.e. SGD, Naïve Bayes, Random Forest and Logistic Regression are used for the preparation and testing of social media abuse content. All four Logistic Regression, Random Forest, Naive Bayes and SGD (Stochastic Gradient Descent) were able to detect the true positives with 77 percent, 93 percent, 75 percent and 79 percent accuracy respectively.

Index Terms— Machine learning, Naïve Bayes, Random Forest, SGD, Cyberbullying.

I. INTRODUCTION

Online bullying is also known as cyberbullying. As the digital realm has grown and technology has progressed, it has become increasingly popular, especially among teenagers. Cyberbullying is when someone, usually an adolescent, on the internet and in other digital spaces, particularly on social networking platforms, bullies or harasses others.

The bullying phrases in the tweet text and tweet responses was identified using machine learning algorithms for cyberbully identification. The output file will be submitted to the classification algorithms after obtaining the output from the preprocessing stage. For identification, a qualified classifier would be used there.

II. RELATED WORK

The popularity of social media has exploded, making it incredibly simple to engage with people from all over the world. As a result, there has been a rise in cyberbullying on social media [1]. Cyberbullying should be discovered and prevented since it can cause substantial emotional and psychological harm. Researchers have put forth a lot of effort to detect cyberbullying in a variety of languages other than Bangla and Romanized Bangla. HaterNet [2], an intelligent technology that recognises and monitors the evolution of hate speech on Twitter and is currently being used by the Spanish nationwide office for Hate Crimes of the Spanish country Secretariat for

Security. The major purpose of this article is to introduce the first intelligent system that uses video display units and visualises, as well as the use of social community.

Social connections evolved within local cultural boundaries [3] such as geo spatial areas prior to the invention of information communication technology (ICT). The recent advancements in communication technologies have significantly surpassed the old communications' time and spatial limits. These social technologies have ushered in a new era of user-generated content, online human networks, and rich data on human behaviour.

The Internet and electronic media have ushered in a new era of communication, and social media platforms are one of the most widely used platforms today [4]. However, some people utilise these sites for harmful purposes, and "cyberbullying" is one of the most common negative characteristics.

Cyberbullying is a type of bullying that takes place through the internet and involves insulting or harming others. Many scholars have presented answers and ways to combat this threat, but sarcasm is one facet that has yet to be addressed.

It provides comprehensive, up-to-date coverage of the expanding issue of cyberbullying [5]. Cyberbullying is a book for counsellors, teachers, school administrators, and those who work with children and teenagers. It discusses the real-life threats that youngsters confront online. Features and Advantages Different types of cyberbullying and cyberbullying contexts are discussed. A look at some of the most popular theories of aggression. The effects of cyberbullying and victimisation on children's development.

Purpose With the growth of social media platforms, there has been an increase in the number of incidences of cyberbullying. Every day, a big number of people, particularly teens, are victims of cyberbullying [6]. A cyberbullied person's psyche can be permanently altered. The victim may develop social anxiety, self-harm, despair, or, in the worst-case scenario, suicide as a result of it. Using machine learning, this research evaluates various strategies for automatically detecting cyberbullying from tweets.

Social media has had a significant global impact on many aspects of human communication and has become a daily activity in both personal and professional lives[8]. This article examines the history and evolution of social media, as well as the role it played in the revolution of business communications.

Bullying has moved from classrooms and backyards to cyberspace in the age of the Internet and electronic devices, and it is now known as Cyberbullying [9]. Youngsters in Arab countries, like children everywhere, are victims of cyberbullying. As a result, concerns about cyberbullying are growing. For the goal of dealing with this scenario, a lot of research is done. While prior research focused on the psychological effects of cyberbullying on the victim, the current study focuses on identification and mitigation of cyberbullying.

Cyberbullying is a well-studied problem; nevertheless, research on the automated detection of cyberbullying on social media using ensemble learning is still in its early stages [11]. The goal of this research was to see if a random forest ensemble learning method could be used to detect cyberbullying.

As the size of Twitter's data grows, so do its users' undesired behaviours. Cyberbullying [13] is one example of such unwelcome behaviour, which can have disastrous repercussions. As a result, it's vital to recognise cyberbullying activity as quickly as possible by examining tweets. The most common methods for detecting cyberbullying are largely stand-alone and consequently time-consuming. The principles of collaborative computing are used in this study to better the detection task.

Megan Meier, a thirteen-year-old Missouri girl suffering from melancholy and low self-esteem[14], was approached by sixteen-year-old Josh Evans on the social media network MySpace in October 2006. ' Meier believed "she finally had a boy who she thought actually thought she was lovely," according to her mother. 2 Meier unexpectedly changed his tune after developing a more positive attitude on life, owing in part to her contacts with Evans. 3 He began sending Meier a "dangerous message."

Due to a misunderstanding of the concept of freedom of speech, the introduction of social media, notably Twitter, has raised a slew of concerns [15]. Cyberbullying is one of these challenges, which is a serious global problem that impacts both individuals and societies. Many attempts have been made in the literature to intervene in, prevent, or alleviate cyberbullying; however, these attempts are impractical because they rely on the victims' interactions. As a result, cyberbullying can be detected without the intervention of a third party.

Many young people's well-being, education, family, and peer relationships are impacted by cyberbullying[16]. Cyberbullying is a group phenomenon most frequent at the transitional years between primary and secondary school, according to a recent study of 548 young Australians. It comes in a variety of shapes and sizes, with the roles of 'bully' and 'victim' overlapping. Despite the devastating emotional consequences of cyberbullying, more than a quarter of victims did not seek help, highlighting the need for additional information and support.

Cyberbullying is when someone is bullied through the use of technology[17]. Despite the fact that it has been a problem for many years, the impact on young people has just recently become more widely recognised. Bullies thrive on social networking sites, and teens and young people who use them are particularly vulnerable to

attacks. We can recognise language patterns used by bullies and their victims using machine learning, and design rules to automatically detect cyberbullying content.

Peer aggression's detrimental consequences on mental health are major public health concerns [18]. The goal of this study was to see if there were any links between cyberbullying and school bullying victimisation and suicide ideation, intentions, or attempts among middle and high school children, and if these links were mediated by depression reports.

Cyberbullying is a growing youth problem [20], according to recent studies. We present a novel automated classification approach that recognises cyberbullying messages without having to fit them into a huge dimensional space in this research. A classifier, on the other hand, due to its overfitting difficulty, cannot produce a restricted convergent answer. In light of these constraints, we created a text classification engine that pre-processes tweets, removes noise and other background information, extracts the desired features, and classifies them without overfitting the data.

III. PROPOSED MODELLING

The classified twitter data is used to train the system and the prediction pipeline is used to contrast the performance of different common classification algorithms and to decide the most fitting model used to classify the real-time tweets as 0 digit for non-cyberbullying and 1 digit for the existence of cyberbullying such as accuracy, precision, recall, f-1 score. The detailed work flow is mentioned in the figure 1.

Economic Feasibility Technically, a system can be constructed, but it must still be a good investment for the business if it is installed. The cost of developing the system is weighed against the overall benefit of the new systems in the economic feasibility analysis. Benefits must be equal to or greater than the costs. Economically, the arrangement is viable. There is no additional hardware or software required.

Technical Feasibility Earlier, there is no valid system available to measure the accuracy and percentage of cyberbullying in the social media. Our project will run the four different machine learning algorithms to detect the accuracy. This project can be easily accessed by the victims, if the project is connected to a web hosting page.

Social Feasibility ahead, the control problems and users requirements were taken into consideration. so, there's no question of resistance from the users that can undermine the feasible applicational blessings. The nicely-planned. layout might ensure the most useful usage of the computer sources and would assist inside the improvement of overall performance reputation.

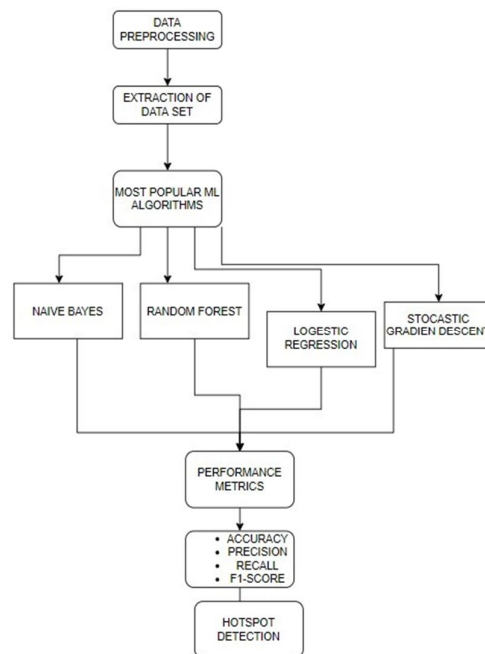


Figure 1. Work Flow of Proposed System

Random forest is a supervised learning method. It produces a "forest" out of a collection of decision trees that are often trained using the "bagging" method. The main idea of the bagging method is that combining many learning models enhances total output. Random forest has the advantage of being able to address classification and regression problems, which are common in machine learning systems nowadays. Random forests were used extensively in this proposed work to classify bullies and non-bullies. It divides the data into subsamples and constructs decision trees for each one. The information is then analyzed based on the outcomes of the tree voting. It uses averaging to improve efficiency while reducing over-fitting.

A classification approach known as logistic regression is used to assign data to a discrete set of classes (LR). Some examples of classification challenges are email spam or not spam, online transaction fraud or not fraud, and tumour malignant or benign. The Logistic Sigmoid Function is used to turn the result of Logistic Regression into a probability value (LSF). A classification procedure called logistic regression is used to assess the likelihood of a reference variable. It is one of the most basic machine learning algorithms that may be used to solve a variety of classification issues, such as spam detection, diabetes prediction, cancer diagnosis, and so on.

SGD — Stochastic gradient descent can be used to optimise an objective function with sufficient smoothness properties. For example, instead of the genuine gradient (taken from the entire data set), an estimate of that gradient is used (calculated from a randomly selected subset of the data). This reduces the computational burden in high-dimensional optimization problems, allowing for faster iterations at the tradeoff of a slower convergence time. Stochastic gradient descent has become a common optimization method in machine learning. The SGD algorithm is a classification algorithm.

At each iteration, it chooses one data point at random from the full data collection, considerably reducing the equations. It's also usual to sample a small number of data points at each stage rather than just one, which is known as 'mini-batch' gradient descent.

A naive Bayes classifier (NBC) uses the Bayes theorem to categorise items. Naive Bayes classifiers assume that data point properties are highly independent. Naive Bayes classifiers are commonly used in spam filters, text analysis, and medical diagnosis. These classifiers are often employed in machine learning because of their simplicity. By obtaining the largest probability value, the Naive Bayes classifier method is used to categorise test data into the most appropriate category. The Tweets document is utilised as test data in this investigation. There are two stages to document categorization.

The first level comprises of practise using papers that have well defined categories. The second stage is the document classification method, which has yet to be assigned a category. It is decided to apply the naive Bayes classification technique.

IV. RESULTS AND DISCUSSIONS

The findings of the experiments are described in this section and their importance is discussed. First, the output outcomes of each classifier were identified and discussed in Table, where it is defined; It displays each classifier's assessments in terms of accuracy, precision, recall and F1 score.

The proposed model utilizes two separate attribute extraction strategies for the chosen four ML classifiers. To obtain greater precision, these methods were set up empirically. In the dataset, for illustration, Random Forest obtained the highest precision and F1 score, where 94 percent and 0.92 respectively are the classification accuracy and F1 score. In the meantime, there is a small difference between the output of LR, SGD, and NB classifiers, where SGD reached 79 percent accuracy, but the F1 value was lower than Random Forest. As seen in Table 1, this means Random Forest does better than other classifiers.

TABLE I. PERFORMANCE RESULT OF CLASSIFIERS

Algorithm	Accuracy	Precision	F1 Score	Recall
Naïve Bayes	75%	0.81	0.61	0.49
Random Forest	94%	0.92	0.92	0.93
Logistic Regression	77%	0.75	0.69	0.64
SGD	79%	0.73	0.73	0.73

Finally, in the dataset, NB has achieved the lowest precision and recall and F1-Score, as seen in Figure 2.

Evaluation In this review, the feasibility of a proposed model was tested by using serval appraisal tests to determine how effectively cyberbullying can be distinguished from non-cyberbullying by the model. Four machine learning algorithms, namely LR RF, Naive Bayes, and SGD, have been constructed in this review. To understand the success of competing models, it is important to review common measurement measures in the research community. The most frequently used parameters for assessing cyberbullying classifier SM platforms are as surveys:



Figure 2: Classification of Algorithms

Accuracy: Precision measures the percentage of the individual cases observed to the total cases, and has been used to test cyberbullying prediction models. It can then be estimated as follows:

$$\text{Accuracy} = (tp + tn) / (tp + fn + tn + fn) \quad (1)$$

tp is true positive, tn is true negative, fp is false positive, and fn is false negative.

Precision: With true positive (tp) and false positive (fp) tweets belonging to a particular category, Precision measures the proportion of related tweets.

$$\text{Precision} = (tp) / (tp + fp) \quad (2)$$

Recall: The ratio of relevant tweets retrieved to the total amount of relevant tweets is determined by recall.

$$\text{Recall} = tp / (tp + fn) \quad (3)$$

F1-Score: It offers a way to mix accuracy and recall into a single metric that captures both properties.

$$\text{F1 - Score} = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall}) \quad (4)$$

V. CONCLUSIONS

Inside western cultures, cyberbullying has been a significant concern. A cyber-bully suggested by this project was model of detection by which different methods of text classification were focused on machine learning been prosecuted. The tests were performed on the kaggle platform on a labeled Twitter dataset. In the dataset, the classification accuracy, F1 score and Recall are 94% , 0.92 and 0.93 respectively, the experimental results show that RF obtained the highest precision and F1 score and Recall.

REFERENCES

- [1] Ahmed, M.T., Rahman, M., Nur, S., Islam, A. and Das, D., 2021, February. Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study. In 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1-10). IEEE.
- [2] Al-Ajlan, M.A. and Ykhlef, M., 2018. Deep learning algorithm for cyberbullying detection. International Journal of Advanced Computer Science and Applications, 9(9), pp.199-205.
- [3] Al-Garadi, M.A., Hussain, M.R., Khan, N., Murtaza, G., Nweke, H.F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H.A. and Gani, A., 2019. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. IEEE Access, 7, pp.70701-70718.

- [4] Ali, A. and Syed, A.M., 2020. Cyberbullying Detection using Machine Learning. *Pakistan Journal of Engineering and Technology*, 3(2), pp.45-50.
- [5] Bauman, S. (2014) *Cyberbullying: What Counselors Need to Know*; John Wiley & Sons: Hoboken, NJ, USA., 3. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and Monitoring Hate Speech in Twitter. *Sensors* 2019, 19, 4654.
- [6] Bharti, S., Yadav, A.K., Kumar, M. and Yadav, D., 2021. Cyberbullying detection from tweets using deep learning. *Kybernetes*.
- [7] Bozyigit, A., Utku, S. and Nasibov, E., 2021. Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, 179, p.115001.
- [8] Edosomwan, S.; Prakashan, S.K.; Kouame, D.; Watson, J.; Seymour, T. (2011) The history of social media and its impact on business. *J. Appl. Manag. Entrep*, 16, 79–91.
- [9] Haidar, B., Chamoun, M. and Serhrouchni, A., 2017, October. Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In *2017 1st Cyber Security in Networking Conference (CSNet)* (pp. 1-8). IEEE.
- [10] Haidar, B., Chamoun, M. and Yamout, F., 2016, November. Cyberbullying detection: A survey on multilingual techniques. In *2016 European Modelling Symposium (EMS)* (pp. 165-171). IEEE.
- [11] Henry, C.C., 2021. Evaluating the Effectiveness of an Ensemble Random Forest Machine Learning Algorithm in Detecting Cyberbullying in the 4chan Politically Incorrect Board Social (Doctoral dissertation, Marymount University).
- [12] Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L., 2013, May. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference* (pp. 195-204).
- [13] Mangaonkar, A., Hayrapetian, A. and Raje, R., 2015, May. Collaborative detection of cyberbullying behavior in Twitter data. In *2015 IEEE international conference on electro/information technology (EIT)* (pp. 611-616). IEEE.
- [14] Miller, K.(2016) *Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law’s limited available redress*. *S. Cal. Interdisc. Law J.*, 26, 379.
- [15] Muneer, A. and Fati, S.M., 2020. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11), p.187.
- [16] Price, M.; Dalglish, J (Aust. 2010). Cyberbullying: Experiences, impacts and coping strategies as described by australian young people. *Youth Stud.*, 29, 51.
- [17] Reynolds, K., Kontostathis, A. and Edwards, L., 2011, December. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops (Vol. 2, pp. 241-244)*. IEEE.
- [18] Sampasa-Kanyinga, H.; Roumeliotis, (2001) P.; Xu, H. Associations between Cyberbullying and School Bullying Victimization and Suicidal Ideation, Plans and Attempts among Canadian Schoolchildren. *PLoS ONE* 2014, 9, e102145.
- [19] Smith, P.K. (2015) *Cyberbullying and Cyber Aggression*. In *Handbook of School Violence and School Safety*; Informa UK Limited: Colchester, UK, issue 2 vol 1.
- [20] Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G. and Rajan, A.R., 2021. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Computers & Electrical Engineering*, 92, p.107186.