iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54329

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

A Monitoring System for Cyber Bullying Activities using Stochastic Gradient Descent Classifier

Payal Budhe[1], Dipalee Rane[2]
[1, 2]*Computer Department, Savitribai Phule Pune University*

Abstract: ***Life has reached a stage where we cannot live without internet enabled technology. New devices and services are being invented continuously with the evolution of new technologies to improve our day-to-day lifestyle. At the same time, this opens many security vulnerabilities. Cybercrime may happen to any device/service at any time with worst ever consequences. Internet use today has a greater impact on young people than ever before. They view the internet and mobile phone networks as the two major communication frameworks that are crucial to our everyday lives and the formation of our identities. However, these technologies are often used improperly. Many internet users are the targets of bullying, which leaves the "target" completely perplexed. Cyberbullying is drastically increasing. The issue of cyberbullying is saddening because the system that enables communication and information flow is evolving into a risky "site" to visit. Cyberbullying affects people all across the world, not just in one nation. United States have begun to enact laws that are focused on cyberbullying. Other countries have adopted laws against bullying that apply to both regular bullying and cyberbullying. The internet gives users the option to browse anonymously and to create profiles with secret identities. Our proposed structure can significantly increase the existing methods detection capacity in actual social network scenarios while effectively making up for their drawbacks.***
Keywords: Cyberbullying, social media, SGD Classifier, Cybercrimes, Technology

I. INTRODUCTION

With more than four billion Internet users globally, the online world has had an enormous impact on society and has become a necessary component of daily life. The current society is entirely dependent on technology, and owing to the internet, young people are now enjoying modern ways of life. One of the major problems resulting from this rapid technological improvement, which also has many drawbacks, is cyberbullying. The internet has grown into a versatile tool that has significantly improved our day-to-day activities. Cyberbullying is only one of many unwanted behaviors that have found their way onto the internet.

A. Cyberbullying

Cyberbullying, also referred to as cyber harassment, is when someone is threatened, bullied, harassed, or scared using specific internet tools. Online bullying is another word for this. Cyberbullying is bullying committed via a digital tool, channel, or platform. Posing as someone else or breaking into someone's account or profile isn't always part of cyberbullying, But there are a lot of different ways that cyberbullying can happen. Cyberbullying is the act of distributing false information about another person online, including through text messages sent by SMS, online chat rooms, game forums, social networking sites, and online chat. It can be viewed on a variety of digital devices, including tablets, smartphones, and laptops. When offensive, harmful, or inappropriate content is sent, uploaded, or shared using various digital tools, it is referred to as cyberbullying. Cyberbullying has become a widespread issue because everyone uses social networking sites today, and it's easy to take advantage of this access. Embarrassing, blackmailing, disparaging, manipulating, or harassing behaviors are included in this. Such hostile behavior readily and unfavorably causes serious harm to a person.

B. Cyberbullying Types

According to the literature, there are 10 types of cyberbullying [34]:

1) Exclusion: Exclusion is the deliberate removal of someone. Exclusion is a factor in both online bullying and physical bullying scenarios where a victim is targeted. For instance, your child can be shut out of message threads or chats involving mutual friends, or they might be refused entry or an invitation to events while they witness other friends receiving one.

2) Harassment: Many forms of cyberbullying fall into the broad category of harassment, but in general, it refers to a consistent pattern of hurtful or threatening online messages made with the purpose of harming someone.

3) Doxing: The act of publicly disclosing private or sensitive information about someone without that person's consent in an effort to embarrass or humiliate them is referred to as outing, also known as doxing. This can include sharing preserved personal conversations in an online private group or disseminating private images or papers of famous people. The victim's lack of permission is crucial.
4) Trickery: Outing and trickery are comparable, with the addition of deception. In these circumstances, the bully will approach their target and deceive them into believing they are safe. Once the bully obtains the target's trust, they take advantage of it by disclosing the victim's secrets and personal information to one or more third parties.
5) Cyberstalking : Cyberstalking is a particularly severe type of online bullying that includes threats of actual physical damage to the victim. It frequently involves offline stalking and may include monitoring, fanciful allegations, threats, and stalking. It is a crime, and the culprit may face a restraining order, probation, or possibly a prison sentence.
6) Fraping : Fraping is when a bully posts offensive stuff using your child's name on social media accounts. When friends publish amusing things on each other's profiles, it can be innocent but also extremely dangerous. For instance, a bully might post homophobic or racial remarks on another person's profile to harm that person's reputation.
7) Disguising : When a bully establishes a fake profile or identity online with the express intention of cyberbullying someone, this is known as masquerading. This can entail choosing a new identity and set of images to deceive the victim, as well as creating a false email account and social media presence. In these situations, the bully is frequently someone the victim knows well.
8) Dissing : Dissing is the act of a bully spreading negative details about their victim through public posts or private messaging in an effort to harm their reputation or interpersonal connections. In these circumstances, the bully frequently knows the victim personally, either as a friend or a mutual friend.
9) Trolling : By making offensive comments online, a bully who wants to disturb others is engaging in trolling. While trolling may not necessarily be considered a form of cyberbullying, it can be when done with malicious and damaging intent. These bullies typically have little personal connection to their victims and are more disengaged from them.
10) Flaming : This kind of cyberbullying consists of posting about or delivering insults and vulgar language to the subject directly. Similar to trolling, flaming usually involves a more direct attack on the target in an effort to instigate an online fight.

C. *Countermeasures By Social Media*

Users can report bullying on social networking sites like Facebook and Twitter, which promote a safe environment online. These include specifying the intended audience, blocking specific users, and recognizing and banning people who behave badly. Despite the fact that they are incredibly important, these techniques are reactive in nature and only apply after the victim has already been harmed. By the time someone reports the offensive post and the required action is taken by the authority, many users may have already read it and experienced the previously mentioned harmful effects. We therefore need an automated system that can rapidly and accurately identify cyberbullying behavior.

D. *Feature types used in cyberbullying prediction*

Table 1. Summary of Content Based feature types used in cyberbullying

	Content Based Features					
Paper	BoW	SG	PF	CB	SF	PR
1	√	√	×	×	×	√
2	√	×	√	√	×	√
3	×	×	√	×	×	√
4	√	×	√	×	×	×
5	√	×	×	×	√	×
6	√	×	√	×	×	×
7	√	×	√	×	√	×
8	√	×	×	×	×	×
9	√	×	√	×	×	×
10	√	×	√	√	√	√
11	√	×	√	√	×	×

12	√	×	×	×	√	×
13	√	×	×	√	×	×
14	√	×	√	×	√	×

BoW - bag of words, SG - skip gram, PF - profanity features, SF - sentiment features, PR – pronouns

Table 2. Summary of Profile Based feature types used in cyberbullying

	Profile Based Features			
Paper	DF	FCF	TSF	LOCF
1	×	×	×	×
2	×	×	×	×
3	√	×	×	×
4	×	×	×	×
5	×	×	×	×
6	×	√	×	×
7	×	×	×	×
8	×	×	×	×
9	×	×	×	×
10	√	√	×	×
11	×	×	×	×
12	×	×	×	×
13	×	×	√	√
14	×	×	×	×

DF - demographic features , FCF - friends or follower count features, TSF - timestamp features, LOCF - location of post feature

1) *Bag of Words:* It is simplifying representation used in natural language processing and information retrieval.
2) *Skip Gram:* It is unsupervised learning technique used to find the most related words for a given word.
3) *Profanity Features:* It should always be used even if only to capture and omit the mist offensive word.
4) *Sentiment Features:* It is the combination an action of belief and emotions that explain for example positive, negative, happy, sad etc.
5) *Pronouns:* A pronoun is defined as a word or phrase that is used as a substitution for a noun.
6) *Demographic Features:* Demographic characteristics are characteristics that describe differences in a society based on gender, age, occupation, level of education, religion, ethnicity, income, marital status and various other aspects of the population.
7) *Friends and Followers count Feature:* The difference between friends and followers is how much access people have to your profile and content. Social media friend is a two-way relationship. When you accept to be someone's friend, you see each other's posts. However , following is a one-way relationship. You see content from the person you follow, but they don't see yours.
8) *Timestamp Features:* A timestamp is a time registered to a file, log, or notification that records when data is added, removed, modified, or transmitted.
9) *Location Post Features:* A location is the place where a particular point or object exists.

II. RELATED WORK

There are several works done on cyberbullying detection.

In [23],This article introduces a brand-new Bully Net architecture for locating bullies on the Twitter social network. In order to create an SN based on bullying tendencies, researchers conducted in-depth research on mining SNs for a better understanding of the interactions between users in social media. They found that by creating conversations focused on environment as well as content, they could successfully pinpoint the feelings and actions that cause bullying. During the experimental investigation, the examination of their suggested centrality metrics to recognize bullies from SN, they were able to identify bullies for a variety of scenarios with about 80% accuracy and 81% precision.

In [18] this research, researchers suggested a detection architecture for cyberbullying to address the issue. They talked about the data architecture for hate speech on Twitter and personal attacks on Wikipedia. Given that tweets containing hate speech typically contained cursing, which made it simple to identify, natural language processing techniques for this type of speech were successful with accuracy rates of over 90% utilizing fundamental machine learning algorithms. Because of this, using BoW and Tf-Idf models rather than Word embeddings models produces better results .Although the three feature selection approaches worked similarly, it was challenging to identify personal assaults using the same model because the comments lacked a lot of learnable sentiment.

In [22], Haider et al.discuss a study on the identification of multilingual cyberbullying. They discovered that the majority of work in this field is done in English, thus they tried to identify cyberbullying in Arabic. They employed ML learning techniques to identify cyberbullying in their work. 32K tweets made up their dataset, and 1800 of those were bullying-related. To identify cyberbullying, they utilized the Support Vector Machine (SVM) and Naïve Bayes methods, and they received F1 scores of 92% and 90%, respectively.

In [20] this study, researchers developed two ensemble-based voting algorithms to identify sentences that are offensive or not. Every ML algorithm and ensemble technique that was used independently has been outperformed by our suggested model. For the twitter extracted dataset, they had the greatest accuracy. The performance of their model will be evaluated in the future using a variety of diverse datasets, as well as some private datasets. Finally, there are many other types of cyberbullying, including harassment, flame, denigration, impersonation, racism, sexism, etc.

In [16] this paper, the issue of detecting cyberbullying was addressed by the sequential hypothesis testing methodology. More specifically, the objective is to choose when to stop extracting and evaluating features from the message and make a decision. Each communication can be classified into one of two classes (i.e., cyberbullying or normal). In order to achieve this, an optimization function was created in terms of the average cost of the classification technique and the cost of features, and the best possible outcome was found.

III. PROPOSED SYSTEM

The detection of cyberbullying involves the following steps:

1) Open the Kaggle repository and load the dataset.

2) Pre-process the dataset by cleaning the text, tokenizing, stemming, lemmatizing, and removing stop words. After text cleaning, linguistic techniques were utilized to examine the pattern of offensive comments.

3) The dataset was then divided into training and test sets. Train different algorithms on the dataset. Utilize the testing dataset and a variety of metrics to evaluate the effectiveness of the algorithms.

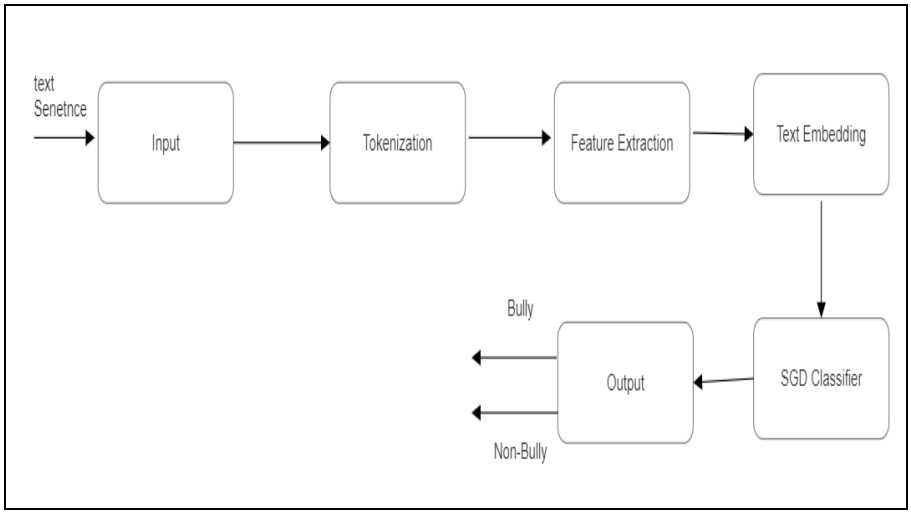


Fig 1. Proposed system

a) Dataset

Gathering data sets from different online networks is the initial stage in the detection of cyberbullying. User comments, posts, pictures, and videos on social networking and media sites typically form data sets for cyberbullying. Using the Twitter API makes it simple to access tweets on Twitter. Along with pre-made datasets from websites like kaggle.com, data from websites like YouTube, Facebook, Myspace, Instagram, etc. are also used for the detection of cyberbullying.

b) Pre-processing

Data pre-processing is the following stage, which is used to modify the data set so that it only contains relevant data. Data pre-processing includes the removal of white spaces, stop words, and special characters prior tokenization and lemmatization. At this stage, we can also use a variety of other methods to organize a data collection.

c) Tokenization

Tokenization is the process of breaking down a piece of text into small units called tokens. A token may be a word, part of a word or just characters like punctuation. Tokenization can be broadly classified into 3 types – word, character, and sub word (n-gram characters) tokenization. Word Tokenization is the most commonly used tokenization algorithm.

d) Stemming

After splitting sentences into words i.e., tokenization humans want to reduce the words to their base or root form. Essentially, this is exactly what is meant by stemming. The process of condensing words with comparable meanings into their "stem" or "root" forms is known as stemming.

e) Lemmatization

Lemmatization is the process of combining a word's several inflected forms into a single unit for evaluation. Similar to stemming, lemmatization adds context to the words. As a result, it links words with related meanings together.

f) Stopword Removal

The most frequent words in any language that have no meaning are called stop words, and natural language processing typically ignores them. Stop words in English include "a," "and," "the," and "of." Stop words are frequently eliminated from texts in natural language processing before they are processed for analysis. This is done to simplify the content and exclude unnecessary information.

g) Feature Extraction

A dimensionality reduction technique called feature extraction divides a large amount of raw data into smaller, easier-to-process groups. These huge data sets share the characteristic of having many variables that demand a lot of computational power to process. The term "feature extraction" refers to techniques for choosing and/or combining variables into features, which significantly reduces the amount of data that needs to be processed while effectively and fully characterizing the initial data set. Text is transformed into a matrix (or vector) of features using feature extraction algorithms. Among the most widely used techniques for feature extraction are: Bag-of-Words and TF-IDF.

h) Text Embedding or Word Embedding

It is an approach for representing words and documents. Word Embedding or Word Vector is a numeric vector input that represents a word in a lower-dimensional space. It allows words with similar meaning to have a similar representation. They can also approximate meaning. A word vector with 50 values can represent 50 unique features.

i) SGD(Stochastic Gradient Descent)Classifier

Gradient Descent is a generic optimization algorithm capable of finding optimal solutions to a wide range of problems. The general idea is to tweak parameters iteratively in order to minimize the cost function. An important parameter of Gradient Descent (GD) is the size of the steps, determined by the learning rate hyperparameters. If the learning rate is too small, then the algorithm will have to go through many iterations to converge, which will take a long time, and if it is too high, we may jump the optimal value. The word ‘stochastic ‘means a system or process linked with a random probability. Hence, in Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. Last step in the detection of cyberbullying ,The information is divided into instances of positive or negative cyberbullying, i.e., information that most definitely contains information about cyberbullying against information that doesn't significantly includes information about cyberbullying. A training collection of labelled examples is required for classification algorithms to predict the label of an input before classifying input data. For data classification, a variety of algorithms and techniques can be utilized.

IV. RESULTS

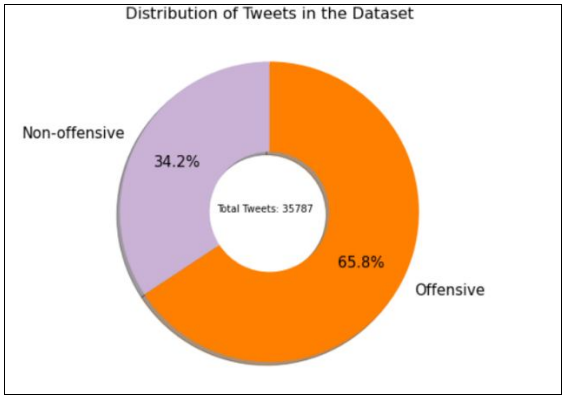


Fig 2. Distribution of tweets in the Dataset.

A. *Performance Metrics*

1) *Accuracy:* The Accuracy measure is the ratio of the number of bully users detected to the total number of bullies. It Doesn't perform well with imbalanced dataset[23] :

Accuracy = # of detected bullies/total number of bullies.

2) *Precision and Recall:* Precision and recall are evaluation metrics used in binary classification tasks. Precision is the measure of exactness and recall is the measure of completeness. They are defined as follows[23] :

Precision = # of true bullies detected/ total number of detected users.

Recall = # of true bullies detected/total number of true bullies

3) *F1 Measure:* F1 measure is the harmonic mean between precision and recall. The range for F1 is [0,1]. It measures how many bullies are identified correctly and how robust it is. Mathematically, it can be expressed as[23] :

F1=2*Precision*Recall/Precision + Recall

Table 3. Performance Evaluation Metrics

Classifiers	Accuracy	Precision	Recall	F1
SGD	92.73%	0.969	0.918	0.943
SVM	89.75%	0.885	0.896	0.886
J 48	89.71%	0.890	0.901	0.886
Naïve Bayes'	75.52%	0.858	0.802	0.791
Random Forest	86.57%	0.898	0.907	0.864
Signed Networks	73.60%	0.813	0.776	0.794

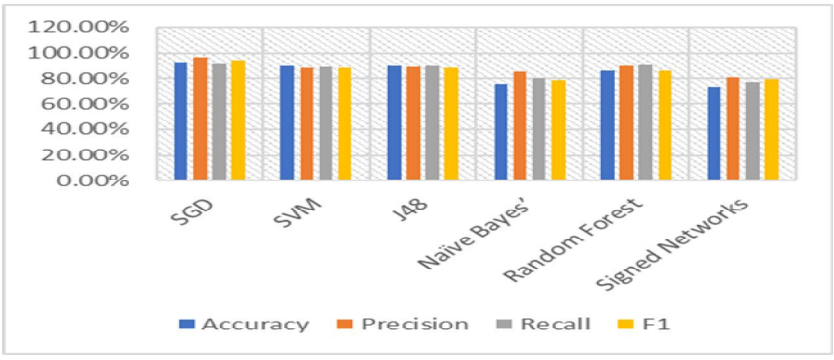


Fig 3. Classification Report

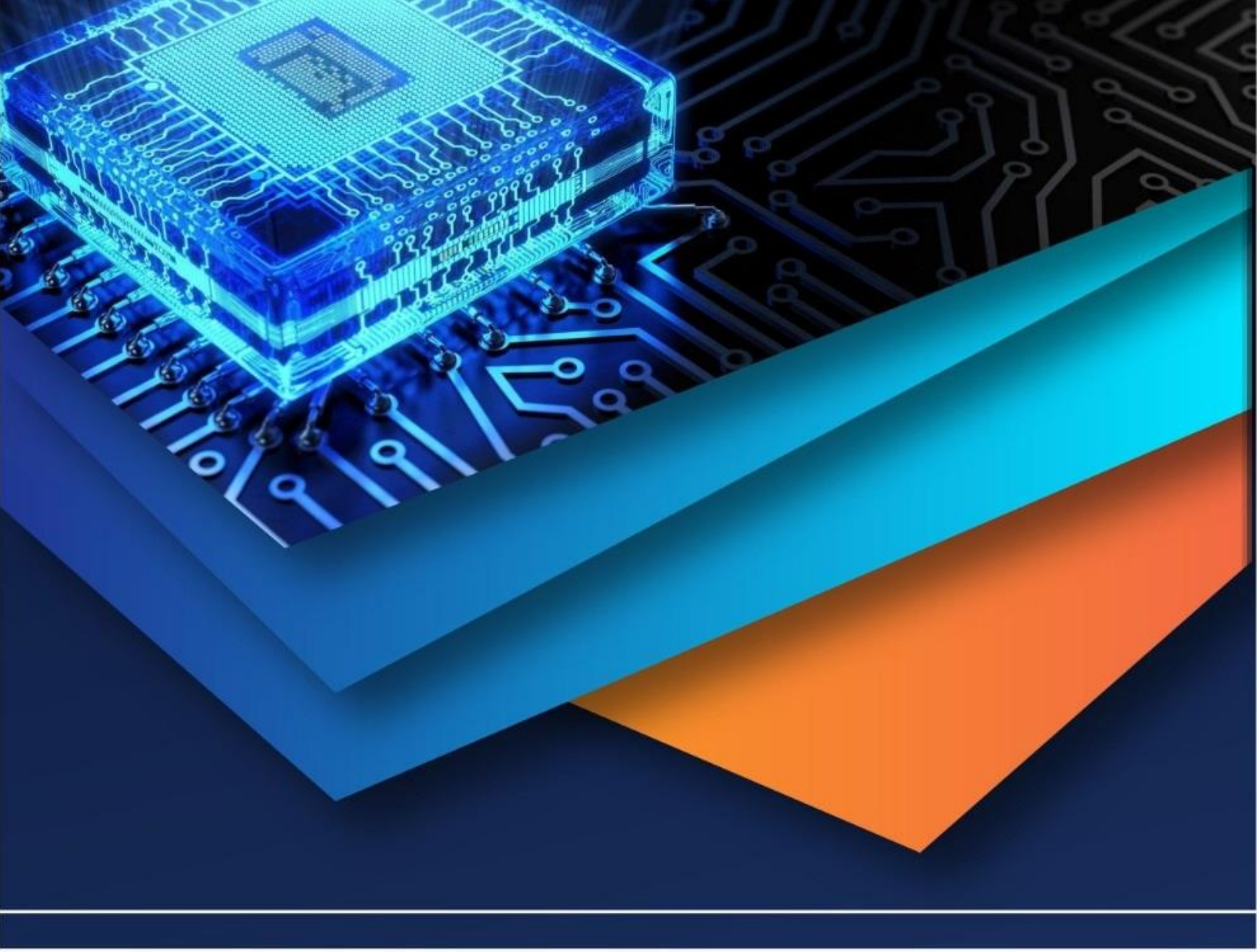
V. CONCLUSION

Cyberbullying has become more common and has begun to generate severe social issues as a result of young people using social media more frequently and the websites that host social media platforms becoming more widely used. A mechanism for automatically identifying cyberbullying must be created in order to stop its harmful effects. Given the significance of identifying cyberbullying, we investigated in this study how to recognize posts on social media that were associated with cyberbullying. This study looked at several studies that investigated the use of different algorithms to identify hostile activity on social networking sites. There was also a list of the numerous discriminatory traits that were used to identify cyberbullying on online social networking sites. With an accuracy of 92.73% and an F-measure of 94.32%, the stochastic gradient descent classifier provides us with the superior outcome. Because of the development of networking and information technology, there are now answers to online contact that are wonderful, awful, hateful, and everything in between. These reactions are routinely mishandled and have left innocent people with lifelong emotional pain, which frequently inspires hopelessness and suicide. They were unable to publicly ask for assistance from various agencies or family members.

REFERENCES

[1] Chavan, V.S. and S. Shylaja. “Machine learning approach for detection of cyber-aggressive comments by peers on social media network. in Advances in computing”, communications and informatics (ICACCI), 2015 International Conference on. 2015. IEEE.
[2] Chen, Y., et al.” Detecting Offensive Language in social media to Protect Adolescent Online Safety. in Privacy, Security”, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom). 2012. IEEE
[3] Dadvar, M., et al., “Improved cyberbullying detection using gender information”. 2012.
[4] Dinakar, K., R. Reichart, and H. Lieberman, “Modeling the detection of Textual Cyberbullying”. 2011.
[5] Van Hee, C., et al. “Detection and fine-grained classification of cyberbullying events”. in International Conference Recent Advances in Natural Language Processing (RANLP). 2015.
[6] Hosseinmardi, H., et al.,” Detection of cyberbullying incidents on the Instagram social network”,arXiv preprint arXiv:1503.03909, 2015
[7] Kontostathis, A., et al. “Detecting cyberbullying: query terms and techniques”. in Proceedings of the 5th annual acm web science conference. 2013. ACM.
[8] Sanchez, H. and S. Kumar, “Twitter bullying detection”. UCSC ISM245 Data Mining course report, 2011.
[9] Zhao, R., A. Zhou, and K. Mao. “Automatic detection of cyberbullying on social networks based on bullying features”, in Proceedings of the 17th International Conference on Distributed Computing and Networking. 2016. ACM.
[10] Squicciarini, A., et al. “Identification and characterization of cyberbullying dynamics in an online social network. in Proceedings of the Advances in Social Networks Analysis and Mining”, ACM-2015.
[11] Reynolds, K., A. Kontostathis, and L. Edwards. “Using machine learning to detect cyberbullying in Machine Learning and Applications and Workshops” (ICMLA), 2011 10th International Conference on. 2011. IEEE.
[12] Yin, D., et al., “Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB”, 2009.
[13] Xu, J.-M., et al. “Learning from bullying traces in social media”, in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012. Association for Computational Linguistics.
[14] Galán-García, P., et al. “Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying”, in International Joint Conference SOCO’13-CISIS’13-ICEUTE’13. 2014. Springer.
[15] Chris Emmery,Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Ve ´ronique Hoste, Walter Daelemans, “Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity”,Springer2020.
[16] Saloni Mahesh Kargutkar, Prof. Vidya Chitre , “A Study of Cyberbullying Detection Using Machine Learning Techniques” , IEEE Xplore 2020.
[17] Mohammed Ali Al-garadi1,Mohammad Rashid Hussain, Nawsher Khan, Ghulam, Murtaza, Henry Friday Nweke1, Ihsan Ali, Ghulam Mujtaba,Haruna Chiroma, Hasan Ali Khattak and Abdullah Gani,” Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges ”, IEEE 2019.
[18] Monirah A., Al-Ajlan, Mourad Ykhlef, “Optimized Twitter Cyberbullying Detection based on Deep Learning”, 978-1-5386-4110-1, IEEE-2018.
[19] N. M. Zainudin, K. H. Zainal, N. A. Hasbullah, N. A. Wahab, and S. Ramli, “A review on cyberbullying in Malaysia from digital forensic “
[20] Vandana Nanda Kumar, Binsu C, Kovoor, Sreeja M.U., “Cyber - Bullying Revelation in Twitter Data using Naïve-Bayes Classifier Algorithm “,International Journal of Advanced Research in Computer Science. Volume 9, No. Jan-Feb 2018.
[21] Semiu Salawu, Yulan He, and Joanna Lumsden, “Approaches to Automated Detection of Cyberbullying: A Survey “, IEEE Transaction 2017.
[22] Rohit Pawar, Rajeev R. Raje, “Multilingual Cyber bullying Detection System” ,IEEE 2019.
[23] Aparna Sankaran Srinath , Hannah Johnson, Gaby G. Dagher , and Min Long, “BullyNet: Unmasking Cyberbullies on Social Networks”,IEEE2021.
[24] Farhan Bashir Shaikh, Mobashar Rehman, and Aamir aamin, “Cyberbullying: A Systematic Literature Review to Identify the Factors Impelling University Students Towards Cyberbullying”, IEEE 2020.
[25] Bandeh Ali Talpur, Declan O’Sullivan, Cyberbullying severity detection: A machine learning approach ,Plos one 2020.
[26] Rekha Sugandhi, Anurag Pande, Siddhant Chawla, Abhishek Agrawal, Husen Bhagat, “Methods for Detection of Cyberbullying: A Survey”, 2015 15th International Conference on ISDA
[27] https://www.bing.com/image/search?q=cyberbullying+detection+diagram & form=HRDSC2&first=1&tsc=ImageHoverTitle
[28] Cyril Onwubiko and Karim Ouazzane,” SOTER: A Playbook for Cybersecurity Incident Management”,IEEE 2022.
[29] Norita Ahmad , Phillip A. Laplante ,Joanna F. Defranco , And Mohamad Kassab ,” A Cybersecurity Educated Community”,IEEE 2022.
[30] Piyush Vyas , Martin Reisslein ,Bhaskar Prasad Rimal, , Gitika Vyas , Ganga Prasad Basyal ,and Prathamesh Muzumdar,” Automated Classification of Societal Sentiments on Twitter with Machine Learning”,IEEE 2022.

[31] Shuwen Wang, Xingquan Zhu, Weiping Ding, and Amir Alipour Yengejeh," Cyberbullying and Cyberviolence Detection: A Triangular User-Activity-Content View",IEEE 2022.
[32] Belal abdullah hezam murshed , jemal abawajy , suresha allappa1, mufeed ahmed naji saif , and hasib daowd esmail al-ariki," DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform",IEEE 2022.
[33] Zhongyuan Jiang , Xianyu Chen , Jianfeng Ma and Philip S. Yu," Rumor Decay: Rumor Dissemination Interruption for Target Recipients in Social Networks",IEEE 2022.
[34] 10 Forms of Cyberbullying | Kids Safety (kaspersky.com).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY