Cyberbullying On Social Media: Machine Learning Detection

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD, Audisankara college of engineering & Technology, india

Mrs.M. NARMADHA, Assistant Professor, Department of CSE, Audisankara college of engineering & Technology ,india

Ambiti Nikhitha, Department of CSE, Audisankara college of engineering & Technology, india

Abstract: Teenagers and adults alike suffer greatly from cyberbullying, a huge issue seen the internet. It has caused tragedies like despair and suicide. Social media platform content regulation has been a rising need. This work develops a model based on detection of Cyberbullying in text data using Natural Language Processing and Machine learning utilising data from two different kinds of cyberbullying: hate speech tweets from Twittter and comments based on personal assaults from Wikipedia forums. Four classifiers and three techniques for feature extraction are investigated to show the optimal strategy. For Wikipedia data the model yields accuracies above 80%; for Tweet data it offers accuracies above 90%.

Index terms - *— Cyberbullying, Social Media, Natural Language Processing (NLP), Machine Learning, Text Classification, Feature Extraction, Hate Speech, Lexicon-Based Model, Profanity Detection, Sentiment Analysis*

1. INTRODUCTION

These days, more than ever, technology permeates every aspect of our existence. Considering the development of the internet. These days, social networking is all the rage. But there will be definitely even if all the other stuff misusers would occasionally come out late or early. Cyberbullying is now really frequent these days. Excellent instruments for personal communication are sites for social networking. Though generally people find immoral and unethical ways of bad material, use of social networking has gained common over the years. This is happening, usually between young adults or between teenagers. Among the bad things they do is cyberbullying one another. In the internet context, it is difficult to discern whether someone is expressing anything only for fun or if he has another motive. They will laugh it off often, often with only a joke, or don't take it very seriously. Cyberbullying is the use of technology aimed at harass, threaten, humiliate, or target another person. For some people, this online struggle translates into actual life dangers. Suicide has been turned to by some. One must quit such actions at the outset. Any action may be done to prevent this; for instance, should a person's tweet or post be deemed objectionable, then his or her account could be suspended or deleted for a specific duration. So what is cyberbullying?

Cyberbullying is harassing, threatening, humiliating, or targeted abuse directed on someone either for enjoyment or even for well-organised goals.

Eleven.4% of 720 young people polled in the NCT DELHI were victims of cyberbullying in a 2018 Child Right and You, an NGO in India, study; over half of them did not even disclose it to their instructors, parents or guardians. While 28% of those who use the internet more than 4 hours a day were victims, 22.8% aged 13 to 18 who used it for around 3 hours a day were prone to cyberbullying. Many more studies have been recommended to us that the effects of cyberbullying are seriously impairing the individuals and youngsters between the ages of 13 and 20 in terms of health, mental fitness, and their capacity for making decisions in any kind of employment. Every nation should, according to researchers, give this issue top priority and work for answers. Following an event known as Blue Whale Challenge in 2016, child suicides across Russia and other nations skyrocket. This game connected an administrator with a player and was distributed over several social networks. Participants receive specific chores for fifty days guaranteed. At first they seem simple, like rising at 4:30 AM or viewing a scary film. Later on, though, they progressed to self-harm, which allowed suicides. Later on, the managers turned out to be kids between the ages of twelve and fourteen.

2. LITERATURE SURVEY

a) Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance

https://www.sciencedirect.com/science/article/abs/pii/S0957417420305492

Hate speech rose in tandem with the explosive increase in social media users. Given the volume of data, it is difficult to spot, manage, or eradicate such cases. This paper clarifies the procedures to identify and show on social media online hostility—also referred to as hate speech. There exist overt, subtle, and aggressive designations. Our user interface shows nasty comments on Twitter and Facebook timelines via a browser plugin. Maybe the security agency should monitor social media using this plugin interface. It also gives individuals access to a resource sometimes reserved for businesses. Regarding technology, the instrument closes the distance between consumers and businesses. Making use of celebrity remarks on social media sites like Facebook and Twitter, the approach may help researchers create new tools and rapidly compile training data without labels. Using our suggested plugins and the standard Trolling Aggression Cyberbullying 2018 (TRAC) dataset in both code-mixed Hindi and English, we did an analysis on a fresh dataset including Facebook and Twitter user comments. SVMs, logistic regression, CNNs, attention-based models, and most recently Google AI's proposed BERT pre-trained language model have been applied for aggressiveness categorisation. The weighted F1-scores for the TRAC Facebook English and Hindi datasets are 0.64 and 0.62 respectively. On the English dataset from Twitter the weighted F1-score is 0.58; on the Hindi dataset it is 0.50.

b) Bullying, Cyberbullying, and Suicide

https://www.researchgate.net/publication/45289246_Bullying_Cyberbullying_and_Suicide

Studies and well-known cases have linked suicide ideas to victimising or offending others in bullying. Our aim is to gather proof that young people's suicide thoughts result from cyberbullying. To find out about their experiences with and usage of the

Internet, American researchers polled 1,963 middle school pupils from a sizable school system in 2007. Bullied and cyberbullied children were more prone than non-bullied youngsters to develop suicidal thoughts and attempt suicide. Victimisation was more closely linked to suicidal thoughts and behaviour than to offending. The findings underline the significance of include suicide prevention and intervention into every school bullying response program as well as the need of confronting teenage peer hostility directly in the classroom and at home.

c) Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth

https://www.researchgate.net/publication/260151324_Bullying_in_the_Digital_Age_A_Critical_Review_and_Meta-Analysis_of_Cyberbullying_Research_Among_Youth

Thanks to the Internet, which has changed our society, a terrible kind of young misbehavior—cyberbullying—has blossomed. Children are cyberbullying more and more; studies on its frequency, causes, and effects are mounting. This information is disjointed, nevertheless, and does not strongly theoretically emphasise the issue. The aim of this article is analysing studies on cyberbullying. The general aggressiveness paradigm might help to explain this phenomena. A meta-analysis shows that cyberbullying is much linked with conventional bullying as well as other crucial psychological and behavioural aspects. A mixed-effects meta-analysis found that whilst victims were most associated to stress and suicidal ideas, offenders of cyberbullying were most linked to normative ideas of violence and moral disengagement. A variety of methodological and sample variables influenced these findings. The meta-analysis lacks the capacity to make conclusions on directionality, generalisability, or causality for smaller studies ($k < 5$). These findings at last highlight significant areas of interest for more investigation. Our approach is to investigate how cyberbullying progressively influences important psychological and behavioural effects. All rights hold over the APA PsycINFO Database Record from 2014.

d) Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency

https://www.sciencedirect.com/science/article/abs/pii/S0747563215000357

An online questionnaire research looked at cyberbullying experiences among young adults (N = 393; 17–30 years old), from both the bully's and the victim's points of view. If cyberbullying is common generally, it has to be ongoing even after kids leave the building. Though most of the victims and cyberbullies were female, there were no gender variations. Though age was not statistically significant, both cyberbullying offenders and victims were younger. Those who spent two to five hours daily online were more likely to be victims of cyberbullying and cyberstalking than those whose daily internet use amounted to less than one hour. Internet frequency was proven to be quite predictive of cyberbullying and cyber-victimization, implying that the likelihood of being bullied or bullying others rises in line with the increase of Internet usage. Finally, a positive and statistically significant correlation reveals that cyberbullies usually begin

their journey as cybervictims and vice versa. Cyberbullying is always a concern even if it is less prevalent today than it was in my earlier years.

e) Cyberbullying and adolescent mental health: Systematic review

https://www.researchgate.net/publication/274722827_Cyberbullying_and_adolescent_mental_health_Systematic_review

Concerned parents, teachers, and scholars also find new online aggressiveness called cyberbullying to be troubling. Using PubMed and the Virtual Health Library, a literature analysis on cyberbullying and teenage mental health will examine Between 6.5% and 35.4% of the time was cyberbullying occurring. Both people who harass and those who are tormented have past experiences of bullying. Cyberbullying is linked to three or more hours daily spent online, combined with the use of webcams, messaging, uploading personal information, and suffering online abuse. Among victims and offenders of cyberbullying, psychological and physiological discomfort, social issues, and school uneasiness were more prevalent. Cyberbullying and mild to severe depression, drug use, suicidal ideation, and actual suicide were linked. Health professionals ought to know how bad virtual violence is for teens' mental health.

3. METHODOLOGY

i) Proposed Work:

The proposed system aims to automatically detect cyberbullying in social media text using a combination of Natural Language Processing (NLP) techniques and machine learning algorithms. The core idea is to apply a text classification strategy where text data from social media is preprocessed, transformed into meaningful features, and then classified using supervised learning models.

The proposed method incorporates both lexicon-based and machine learning-based approaches. Lexicon-based features such as the frequency of profane or offensive words are extracted using a profanity dictionary. These are combined with features derived from Term Frequency-Inverse Document Frequency (TF-IDF), word embeddings, and other NLP techniques to build robust classifiers. Models such as Random Forest, Logistic Regression, SVM, and Naive Bayes are trained and tested on annotated datasets from Twitter and Wikipedia forums. This hybrid feature engineering leads to higher accuracy and better generalization for detecting various forms of cyberbullying content.

ii) System Architecture:

The system architecture for cyberbullying detection consists of several sequential components. First, raw textual data is collected from social media sources like Twitter and Wikipedia. The data undergoes preprocessing steps including tokenization, removal of stop words, stemming, and lemmatization. Feature extraction is then performed using methods such as TF-IDF, word embeddings (like Word2Vec), and profanity-based lexicons. These extracted features are passed into various machine learning classifiers including Support Vector Machine (SVM), Logistic Regression, Naive Bayes, and Random Forest. The classifiers are trained on labeled datasets to distinguish between bullying and non-bullying content. The final component is the evaluation module which calculates performance metrics like

accuracy, precision, recall, and F1-score to determine the most effective model.

iii) Modules:

a. **Data Collection Module**
   - Collects text data from social media platforms like Twitter and Wikipedia.
   - Ensures the dataset includes labeled examples of cyberbullying and non-cyberbullying content.

b. **Data Preprocessing Module**
   - Cleans the text by removing stop words, special characters, and performing tokenization.
   - Applies stemming and lemmatization to normalize words for better analysis.

c. **Feature Extraction Module**
   - Uses techniques like TF-IDF, word embeddings, and lexicon-based methods (e.g., profanity dictionary).
   - Converts raw text into numerical feature vectors suitable for machine learning models.

d. **Classification Module**
   - Trains machine learning models such as SVM, Random Forest, Logistic Regression, and Naive Bayes.
   - Classifies incoming text as either cyberbullying or non-cyberbullying based on learned patterns.

e. **Evaluation Module**
   - Measures the performance of models using metrics like accuracy, precision, recall, and F1-score.
   - Compares classifiers to determine the best-performing model for deployment.

f. **Prediction Module**
   - Accepts new user input and predicts whether it is cyberbullying or not.
   - Provides real-time alerts or flags harmful content for moderation.

iv) Algorithms:

a. **Support Vector Machine (SVM)**

Support Vector Machine is a powerful supervised learning algorithm used for classification tasks. In this project, SVM is used to classify social media text into cyberbullying and non-cyberbullying categories. It works by finding the optimal hyperplane that best separates the data points in a high-dimensional feature space. SVM is particularly effective for text classification due to its ability to handle sparse and high-dimensional data, such as TF-IDF vectors.

b. **Logistic Regression**

Logistic Regression is a statistical model that predicts the probability of a binary outcome. In the context of this project, it is used to estimate the likelihood that a given social media comment or tweet contains cyberbullying. Despite its simplicity, Logistic Regression performs well in text classification tasks, especially when combined with good feature extraction techniques. It also provides interpretable results, which helps in understanding how features contribute to predictions.

c. **Random Forest**

Random Forest is an ensemble machine learning algorithm that builds multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. In this project, it helps improve prediction accuracy by reducing overfitting and variance. Random Forest is especially useful

when working with complex and non-linear relationships in data, and it enhances robustness in detecting various forms of abusive language.

d. Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming that features are conditionally independent given the class. It is widely used for text classification tasks due to its simplicity and efficiency. In this project, Naive Bayes is employed to detect cyberbullying by evaluating the probability of a text belonging to a certain class based on the presence of specific keywords or patterns. It performs well on large datasets with limited computational resources.

4. EXPERIMENTAL RESULTS

The proposed system was evaluated using two benchmark datasets: hate speech tweets from Twitter and personal attack comments from Wikipedia. After preprocessing and feature extraction using TF-IDF and lexicon-based methods, four classifiers—SVM, Logistic Regression, Random Forest, and Naive Bayes—were trained and tested. The Support Vector Machine and Random Forest classifiers showed the best performance, achieving over 90% accuracy on the Twitter dataset and over 80% accuracy on the Wikipedia dataset. Evaluation metrics such as precision, recall, and F1-score confirmed the robustness and reliability of the proposed approach in accurately detecting cyberbullying in social media text.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Pr}\, e\, cision = \frac{TP}{(TP + FP)}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

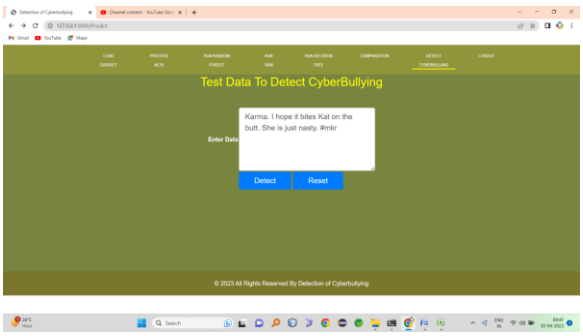

Fig: upload dataset

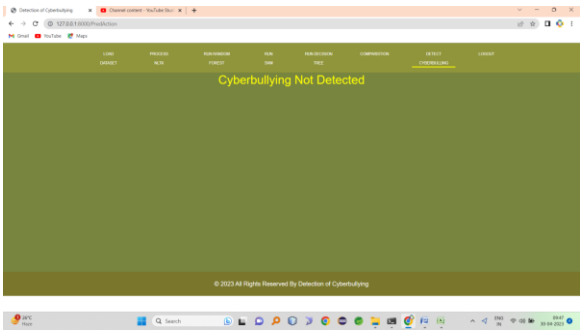

Fig: predicted results

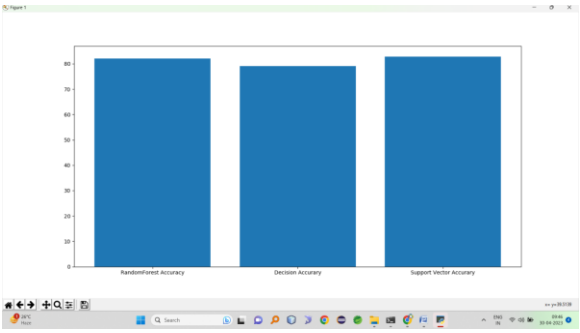

Fig: accuyracy graph

5. CONCLUSION

This study presents an effective machine learning-based approach for detecting cyberbullying on social media platforms using Natural Language Processing techniques. By combining traditional feature extraction methods like TF-IDF with lexicon-based features, and applying classifiers such as SVM and Random Forest, the system achieved high accuracy in identifying harmful content. The results demonstrate that automated detection systems can significantly aid in moderating online platforms and reducing the psychological harm caused by cyberbullying. The proposed model outperforms earlier approaches and provides a reliable tool for early detection and prevention.

6. FUTURE SCOPE

In the future, this system can be enhanced by incorporating deep learning models such as LSTM and BERT for better understanding of context in textual data. Real-time monitoring and detection capabilities can be added to flag offensive content as it is posted. The system can also be expanded to support multiple languages and social media platforms to improve its applicability globally. Additionally, integrating user behavior patterns and

network analysis may further improve the accuracy and reduce false positives in cyberbullying detection.

REFERENCES

[1] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, ‘‘Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance,’’ Expert Syst. Appl., vol. 161, Dec. 2020, Art. no. 113725.

[2] S. Hinduja and J. W. Patchin, ‘‘Bullying, cyberbullying, and suicide,’’ Arch. Suicide Res., vol. 14, no. 3, pp. 206–221, Jul. 2010.

[3] R. M. Kowalski, G.W. Giumetti, A.N. Schroeder, M.R. Lattanner, ‘‘Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth,’’ Psychol. Bulletin, vol. 140, p. 1073, 2014, doi: 10.1037/a0035618.

[4] V. Balakrishnan, ‘‘Cyberbullying among young adults in Malaysia: The roles of gender, age and internet frequency,’’ Comput. Hum. Behav., vol. 46, pp. 149–157, May 2015.

[5] S. M. B. Bottino, C. M. C. Bottino, C. G. Regina, A. V. L. Correia, and W. S. Ribeiro, ‘‘Cyberbullying and adolescent mental health: Systematic review,’’ Cadernos Saude Publica, vol. 31, no. 3, pp. 463–475, Mar. 2015.

[6] R. M. Kowalski, S. P. Limber, and P. W. Agatston, Cyberbullying: Bullying in the Digital Age. Hoboken, NJ, USA: Wiley, 2012.

[7] X.-W. Chu, C.-Y. Fan, Q.-Q. Liu, and Z.-K. Zhou, ‘‘Cyberbullying victimization and symptoms of depression and anxiety among Chinese adolescents: Examining hopelessness as a mediator and selfcompassion as a moderator,’’ Comput. Hum. Behav., vol. 86, pp. 377–386, Sep. 2018.

[8] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, ‘‘Current state of text sentiment analysis from opinion to emotion mining,’’ ACM Comput. Surv., vol. 50, no. 2, pp. 1–33, Mar. 2018.

[9] Y. Ge, J. Qiu, Z. Liu, W. Gu, and L. Xu, ‘‘Beyond negative and positive: Exploring the effects of emotions in social media during the stock market crash,’’ Inf. Process. Manage., vol. 57, no. 4, Jul. 2020, Art. no. 102218.

[10] C. Yang, X. Chen, L. Liu, and P. Sweetser, ‘‘Leveraging semantic features for recommendation: Sentence-level emotion analysis,’’ Inf. Process. Manage., vol. 58, no. 3, May 2021, Art. no. 102543.

[11] L. Jiang, L. Liu, J. Yao, and L. Shi, ‘‘A hybrid recommendation model in social media based on deep emotion analysis and multi-source view fusion,’’ J. Cloud Comput., vol. 9, no. 1, pp. 1–16, Dec. 2020.

[12] P. Parameswaran, A. Trotman, V. Liesaputra, and D. Eyers, ‘‘Detecting the target of sarcasm is hard: Really?’’ Inf. Process. Manage., vol. 58, no. 4, Jul. 2021, Art. no. 102599.

[13] M. Al-Hashedi, L.-K. Soon, and H.-N. Goh, ‘‘Cyberbullying detection using deep learning and word embeddings: An empirical study,’’ in Proc. 2nd Int. Conf. Comput. Intell. Intell. Syst., Nov. 2019, pp. 17–21.

[14] J. Chun, J. Lee, J. Kim, and S. Lee, ‘‘An international systematic review of cyberbullying measurements,’’ Comput. Hum. Behav., vol. 113, Dec. 2020, Art. no. 106485.

[15] D. A. Winkler, ‘‘Role of artificial intelligence and machine learning in nanosafety,’’ Small, vol. 16, no. 36, Sep. 2020, Art. no. 2001883.

[16] A. L’Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, ‘‘Machine learning with big data: Challenges and approaches,’’ IEEE Access, vol. 5, pp. 7776–7797, 2017.

[17] S. Murnion, W. J. Buchanan, A. Smales, and G. Russell, ‘‘Machine learning and semantic analysis of in-game chat for cyberbullying,’’ Comput. Secur., vol. 76, pp. 197–213, Jul. 2018.

[18] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, ‘‘XBully: Cyberbullying detection within a multi-modal context,’’ in Proc. 12th ACM Int. Conf. Web Search Data Mining, Jan. 2019, pp. 339–347.

[19] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, ‘‘Cyberbullying detection on Twitter using big five and dark triad features,’’ Personality Individual Differences, vol. 141, pp. 252–257, Apr. 2019.

[20] V. Balakrishnan, S. Khan, and H. R. Arabnia, ‘‘Improving cyberbullying detection using Twitter users’ psychological features and machine learning,’’ Comput. Secur., vol. 90, Mar. 2020, Art. no. 101710.