ISSN No:-2456-2165

Cyberbullying Detection on Twitter using Machine Learning: A Review

Salisu Suleiman
M.Tech (CSE) Student 4th Semester
Dept. of Computer Science and Engineering
Alakh Prakash Goyal University, Shimla, India

Prashansa Taneja
Assistant Professor
Dept. of Computer Science and Engineering
Alakh Prakash Goyal University, Shimla, India

Ayushi Nainwal
Assistant Professor
Dept. of Computer Science and Engineering
Alakh Prakash Goyal University, Shimla, India

Abstract:- The internet has pervaded every part of human life making it easier to link individuals across the globe and disseminate information to a large group of people. Despite its importance, the cyberworld has a number of negative effects on people today. One of the most dangerous threats in the cyberworld is cyberbullying as it destroys individuals' reputation or privacy, threatens or harasses them, and has a long-term impact on the victim. Despite the issue has been in existence for many years, the impact on young people has just recently become more widely recognized. Using machine learning and natural language processing, the bullies' harassing tweets or offensive comments may be automatically identified and detected. This paper reviewed the previous research in cyberbullying detection domain and more importantly, proposed a novel cyberbullying detection model to close the gap that was discovered during the review of the related literature. In this study, we employed standard supervised learning method and ensemble supervised learning method. The traditional methods used three ML classifiers: Gaussian Naïve Bayes (GNV), Logistic Regression (LR), and Decision Tree (DT) classifiers, While Adaboost and Random Forest (RF) classifiers were used as ensemble technique. We trained and tested our model to detect and classify bullying content as either bullying or non-bullying (binary classification model) using our dataset, and Termed Frequency Inverse Document Frequency (Tf-idf) was used to extract features from a twitter dataset downloaded from kaagle.

***Keywords**:- Cyberworld, Social Media, Machine Learning, Cyber Bullying.*

I. INTRODUCTION

Social media can be termed as a platform that allows users to share anything they want, like; images, documents and videos as well as communicate with others [1]. People engage with one another on social media sites via computers or cellphones. Facebook, Twitter, Instagram, TikTok, and other social media platforms are among the most widely used worldwide [2]. With the rapid expansion of the internet and technology, we have become increasingly dependent on social media in our daily lives. It lets users to exchange information with one another with just a few taps and/or clicks using different applications [2]. It also provides people with entertainment. People have started feeling more sociable despite their current situation, even if they are at home or at work. The social media platforms are widely accessible with smartphones and tablets, and the number of users has increased in recent years.

According to [3] there are nearly four billion Internet users, three billion social media users, and five billion mobile phone users. Despite its importance, social media comes with a slew of problems and challenges. For example, many antisocial behaviors on social media may include; cyberbullying, cyberstalking, and cyberharassment. These behaviours have now become inculcated in our culture and are no longer limited to youth; anybody can be affected.

According to a study taken by [4], cyberbullying affects nearly half of all youths in America. The victim of bullying suffers both physically and mentally due to the harmful nature of the bullying. Because the misery of cyberbullying is too great to bear, victims of cyberbullying commit self-destructive acts like suicide.

Consequently, detecting and combating cyberbullying is critical for teens' safety.

In this era of web 4.0, where people live on digital and online platforms, it is highly difficult to protect society from the frightening rise of cybercrime.

Some examples of cyberbullying attacks are as follows: (1) Sending or posting harsh or abusive comments with the purpose of harming an individual's character (2) Sharing an improper image or video. (3) Creating fake or inappropriate website. (4) Making online threats that lead to someone hurting themselves or harming someone else. (5) Posting hates comments or videos online to promote religious, racial, ethnic, or political hatred. (6) Creating a fake online identity so as to acquire information.

This paper overviewed the contributions, dataset used, results and research gaps of different researches in the topic of cyberbullying.

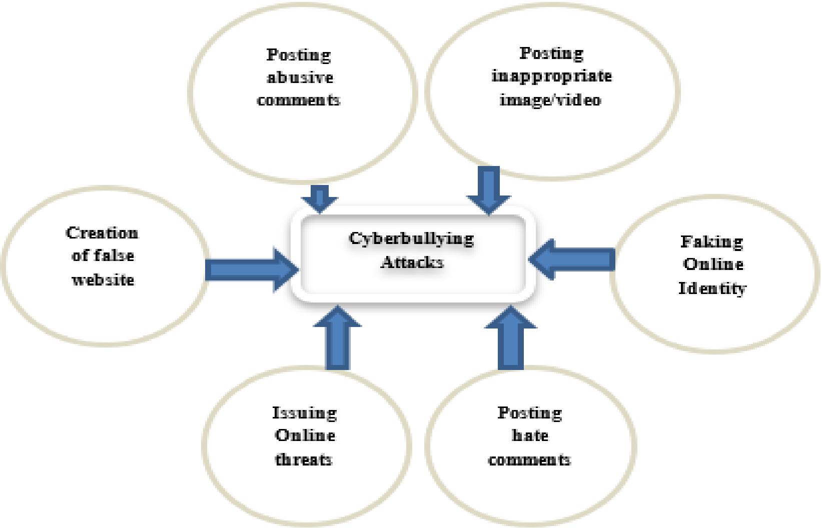

Fig 1: Classification chart for cyberbullying attacks

II. LITERATURE SURVEY

This chapter discusses the problems of cyberbullying and overviewed some previous researches carried out in the field of cyberbullying detection.

According to [1], Four ML algorithms like: SVM, RF, NV and DT were used to identify abusive and bullying messages on social media in English using two features i.e; TF-IDF and BoW to analyse the level of accuracy of four ML algorithms used. Facebook and twitter dataset were successfully downloaded from kaggle.com. The result indicated that SVM outperforms all other machine learning algorithms used in the research. In the same way, TF-IDF outperformed BoW.

In another study, [2] used Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Random Forest (RF), AdaBoost (ADB), Support Vector Machine (SVM) and Naïve Bayes (NB) as machine learning classifiers. Each of these techniques was tested using performance measures such as accuracy, precision, recall, and F1 score to determine the recognition rates of the classifiers when applied to the entire dataset. A global dataset of 37,373 was used to test the seven classifiers used in the study. Logistic regression got the greatest F1-score of 0.928, SGD had the most precision of 0.968, and SVM had the most recall among the classifiers of 1.00. Finally, with a median accuracy of 90.57 percent, the studies demonstrated that Logistic Regression is the best.

According to [5], most of cyberbullying detection studies were conducted in a single language. Because of the detrimental implications of cyberbullying, a model capable of identifying cyberbullying in many languages, including Hindi and Marathi is developed. The dataset was gathered from a number of different sources including newspaper reviews, tourist reviews collected manually, and tweets retrieved from the Twitter API. The result revealed that F1-score has up to 97% and accuracy was measured to be 96%. The percentages were obtained in both Hindi and Marathi. Likewise, Logistics Regression (LR) outperforms SGD and MNB in all the three different dataset used.

In a research done by [6], four machine learning approaches were used to detect bullying text in English, including SVM, LR, RF, and Multilayered perceptron algorithms together with three distinct textual features, Word2Vec, BoW and TF-IDF and a dataset taken from Wikipedia and Twitter. The results indicated that using the same machine learning classifiers, the twitter dataset achieved over 90% accuracy while the Wikipedia dataset achieved over 80% accuracy. The Tf-Idf and BoW features considerably outperformed the Word2Vec function.

According to [7] it is necessary to identify and detect cyberbullying on many social media flat-forms, hence several machine learning techniques such as SVM and Naive Bayes are employed to recognize the presence of bullying messages on Twitter and Wikipedia social flat forms in both Arabic and English languages. In both the twitter and wikipedia datasets utilized, NV outperformed SVM with 90.8 percent accuracy.

Authors in [8] developed a ML technique that uses numerous textual elements to detect cyberbullying on Twitter. In their research, They were successful in creating a series of machine learning models, including linear, tree-based, and deep learning models, the best of which scored above 90% on the four criteria; accuracy, precision, recall, and F-measure.

Authors in [9] used Python and Tensor-Flow to implement their cyberbullying model. They compared DNN's performance to that of standard machine learning models, and find out that DNN-based models are more adaptable to new datasets and outperform the standard machine learning models on the Twitter dataset.

In another study, [10] used a powerful ML classifier namely; Support Vector Machines to detect cyberbullying in English and Dutch languages (SVM). They used LSVM machines in their research, which leverage a large feature set and explore which information sources contribute the most to the task. For the two languages; English and Dutch, the classifier produces F1 scores of 64 percent and 61 percent, respectively.

SVM, RBF, MLP, LR, and SGD algorithms were popular classifiers employed by [11] to develop a model that categorizes comments in datasets according to whether they have cyberbullying or not. To reduce classification time, the classifiers' performance is evaluated using Chi2, Support Vector Machine-Recursive Feature Elimination (SVM-RFE), Minimum Redundancy Maximum Relevance (MRMR), and Relief feature selection techniques. After using feature selection algorithms, the classification times for YouTube, Formspring.me, and Myspace datasets were reduced by 20 times, 2.5 times, and 10 times, respectively. The results revealed that, classifiers with an F-measure value above 0.930, such as Stochastic Gradient Descent (SGD) and Multilayer Perceptron (MLP), outperformed other classifiers, and the SVM-RFE algorithm, which uses the selected 500 features delivered excellent results.

To address cyberbullying issues, the authors in [12] suggested a convolutional neural network cyberbullying detection (CNN-CB) model. They used a Twitter streaming API and obtained their Twitter dataset used in the study, and the results indicated that the CNN-CB algorithm outperformed traditional content-based cyberbullying detection in all the three performance evaluation measures with a 95 percent accuracy.

In [13], it was understood that, according to an Indonesian ministry of communication and Information survey, 58 percent of 435 teenager lack primary understanding about cyberbullying and its negative consequences, and they may have even been the bullies or victims without their knowledge. As a result, a model that can detect cyberbullying actors based on texts and user credibility analysis is needed, as well as alerts them of the consequences of their actions. 257 Healthy Users, 45 Dangerous Bullying Actors, 53 Aggressive Actors, and 6 Potential Bullying Actors were discovered after the users' reliability was appraised. On the twitter dataset, SVM and KNN ML algorithms were applied to detect cyberbullying. The results indicated that SVM performed better and achieved around 67% F1-score.

Another study on cyberbullying identification on social media posts and comments written in Turkish was published by [14]. On the dataset obtained from twitter and Instagram postings, support vector machines (SVM), decision trees (C4.5), Naïve Bayes Multinomial, and K-Nearest To identify cyberbullying, Naïve Bayes Multinomial and K-Nearest Neighbors (KNN) classifiers were used.

When feature selection is used, the classification accuracy for the dataset improves by up to 84 percent, and the Naïve Bayes Multinomial was found to be the most successful in terms of both classification accuracy and running time.

[15] used Linear Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Perceptron, and Logistic Regression algorithms to identify cyberbullying in Japanese writings. The dataset was obtained from twitter and shows a fairly balanced distribution. The Gradient Boosting approach produces the greatest results, with an F1 Score of up to 93.5 percent.

In a study conducted by [16], PHP and HTML, as well as MySQL and the Twitter API, were used to develop a cyberbullying detection model on Twitter. The proposed system will not only be able to recognize a cyberbullying incident, but it will also allow users to identify the identities and information of cyberbullying users (both bullies and victims), as well as conduct location analysis and generate police reports. To discover the F-measure values, J48, NBM, SVM, and IBk classifiers were applied to each fold for both datasets, and then the average of the f-measure values for the five folds was taken. After that, the chi2 and IG feature selection methods are used.

[17] in their study on the identification of cyberbullying on social media, developed a supervised machine learning method for detecting and combating cyberbullying. They employed machine learning classifiers to recognize bullies' linguistic patterns, and then used SVM and NN classifiers to develop a model to automatically detect and prevent cyberbullying. The results showed that, the Neural Network performs better, with an accuracy of 92.8%, while the SVM has an accuracy of 90.3%. On the same dataset provided from kaggle.com, the Neural Network achieves an average f-score of 91.9%, while the SVM achieves an average f1-score of 89.8%.

III. PROPOSED SYSTEM

The major goal of the proposed system is to develop an effective cyberbullying detection model using machine learning algorithms and natural language processing toolkit to deal with the problem of cyberbullying attacks on numerous social media platforms. In this work, we employed two types of supervised machine learning methods: traditional supervised learning methods and ensemble supervised learning methods. The traditional learning method used Gaussian Naïve Bayes, Logistic Regression and Decision Tree classifiers whereas the ensemble methods used Adaboost and Random Forest classifiers. The Dataturks' Tweet Dataset for Cybertroll Detection obtained from Kaggle [18] containing about 20001 instances was used for this study. The content on this page comes from various internet users in the form of tweets and comments. The next stage is data preprocessing, which means we'll prepare our data before putting it into our machine. We start by removing any irrelevant data from our dataset, next we deal with outliers, and finally, we work on any missing data that may be present in our dataset. We used 80% of the dataset to train our model and 20% of the dataset to test it. The model will learn how to classify data into bullying or non-bullying tweets/comments using the training dataset. The testing dataset will be used to evaluate the accuracy of our machine learning model once the machine has been trained.

For performance evaluation, the accuracy of the classifier based on precision, recall, and f1-score of bullying and non-bullying tweets was calculated. The performance evaluation metrics used to assess the quality of the classifier's output are precision, recall and F1-score. Precision is a metric that measures how relevant the results are, whereas recall measures how many relevant results are returned and F1-score is computed as the precision and recall average.

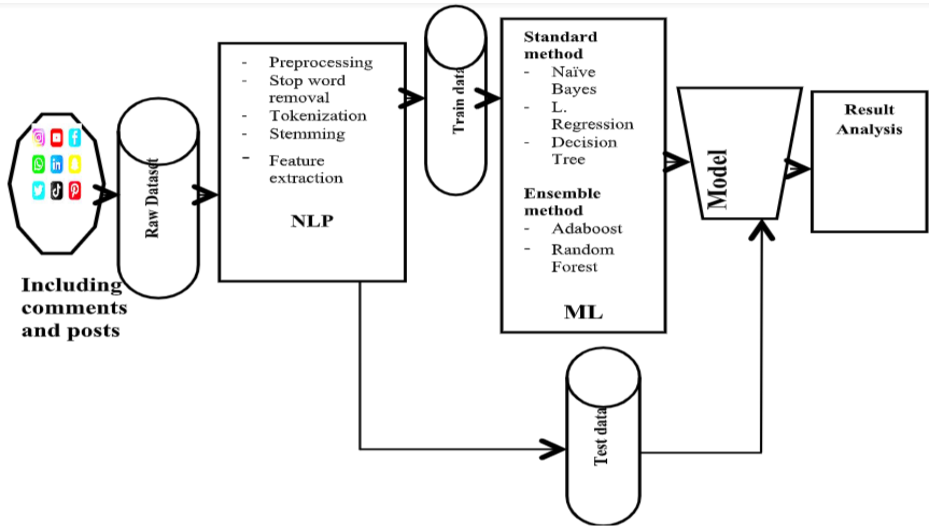

Fig 2: Proposed framework for cyberbullying detection

IV. CONCLUSION AND FUTURE WORK

According to the outcome of our survey, conventional machine learning algorithms are incapable of processing the huge amounts of data generated in Web 4.0, and cyberbullying content cannot be detected effectively. Recently, many researchers are interested in deep learning approaches such as deep recurrent neural networks, Convolutional Neural Network and stacking auto-encoder. The use of these deep learning techniques for precise identification of cyberbullying in social media could be the direction for future research. Another feature work is that, our cyberbullying detection model is binary classification based (bullying or non-bullying), so multi-class classification approach could be also the direction of our future research.

REFERENCES

[1]. M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, “Cyberbullying detection on social networks using machine learning approaches,” *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020.

[2]. A. Muneer and S. M. Fati, “A comparative analysis of Machine Learning Techniques for cyberbullying detection on Twitter,” *Future Internet*, vol. 12, no. 11, p. 187, 2020.

[3]. S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. DeSmet, and I. De Bourdeaudhuij, “Cyberbullying on social network sites. an experimental study into bystanders’ behavioural intentions to help the victim or reinforce the bully,” *Computers in Human Behavior*, vol. 31, pp. 259–271, 2014.

[4]. S. Kemp, “Digital 2019: Global Digital Overview - DataReportal – global digital insights,” *DataReportal*, 13-Apr-2019. [Online].Available:https://datareportal.com/reports/digital-2019-global-digital-overview. [Accessed: 16-May-2022].

[5]. R. Pawar and R. R. Raje, “Multilingual cyberbullying detection system,” *2019 IEEE International Conference on Electro Information Technology (EIT)*, 2019.

[6]. V. Jain, V. Kumar, V. Pal, and D. K. Vishwakarma, *Detection of cyberbullying on social media using machine learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*,2021.

[7]. B. Haidar, M. Chamoun, and A. Serhrouchni, “Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content,” *2017 1st Cyber Security in Networking Conference (CSNet)*, 2017.

[8]. Alduailej, A. H., & Khan, M. B. (2017, September). The challenge of cyberbullying and its automatic detection in Arabic text. In *2017 International Conference on Computer and Applications (ICCA)* (pp. 389-394). IEEE.

[9]. Kargutkar, S., & Chitre, V. Implementation of Cyberbullying Detection using Machine Learning Techniques.

[10]. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, *13*(10), e0203794.

[11]. Çiğdem, A. C. I., Çürük, E., & Eşsiz, E. S. (2019). Automatic detection of cyberbullying in Formspring. me, Myspace and Youtube social networks. *Turkish Journal of Engineering*, *3*(4), 168-178.

[12]. Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep learning algorithm for cyberbullying detection. *International Journal of Advanced Computer Science and Applications*, *9*(9), 199-205.

[13]. H. Nurrahmi and D. Nurjanah, “Indonesian twitter cyberbullying detection using text classification and user credibility,” *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018.

[14]. S. A. Ozel, E. Sarac, S. Akdemir, and H. Aksu, “Detection of cyberbullying on social media messages in Turkish,” *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017.

[15]. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2020). Deep Learning based Recommender System. *ACM Computing Surveys*, *52*(1), 1–38. https://doi.org/10.1145/3285029.

[16]. Hon, L., & Varathan, K. (2015). Cyberbullying detection system on twitter. *IJABM*, *1*(1), 1-11.

[17]. Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social media cyberbullying detection using machine learning. *International Journal of Advanced Computer Science and Applications*, *10*(5), 703-707.

[18]. DataTurks, “Tweets dataset for detection of cyber-trolls,” *Kaggle*, 12-Jul-2018. [Online]. Available: https://www.kaggle.com/dataturks/dataset-for-detection-of-cybertrolls. [Accessed: 16-May-2022].