2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) | 978-1-6654-0637-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IC4ME253898.2021.9768450

A Framework to Detect and Prevent Cyberbullying from Social Media by Exploring Machine Learning Algorithms

Shutonu Mitra, Tasfia Tasnim, Md. Arr Rafi Islam, Nafiz Imtiaz Khan and Mohammad Shahjahan Majib
Department of Computer Science and Engineering, Military Institute of Science and Technology,
Mirpur Cantonment, Dhaka-1216, Bangladesh
Email: shutonumitra@gmail.com, tashfi20@gmail.com, arrrafi24@gmail.com, imtiznafiz@gmail.com, smajib@yahoo.com

***Abstract*—Social media is the most popular way to meet new people and interact with friends and associates nowadays. But unfortunately, users get subject to bully or harassment while surfing through social media. Over the last decade, cyberbullying surfaced as one of the most significant issues in the digital world. Although significant research has been carried out to identify cyberbullying through text-mining techniques on many online platforms, still there is a long way to have a concrete solution to remove cyberbullying from social media. This paper introduces a way for the prevention of cyberbullying from social media by identification of cyberbullying texts (Twitter only) through sentiment analysis, and also classification of cyberbullying according to bullying characteristics depending on the proposed taxonomy. In this context, a suitable framework consisting of three modules (e.g., user interaction, analytics, and decision making) is proposed to prevent cyberbullying from social media. The user interaction module contains user profiles from where posts and comments are taken to the analytics module, the analytics module generates results according to the type of bully and the decision-making module takes action finally. Temporary/permanent ban on posting or commenting, bully badge shown at the personal profile are the actions proposed. However, in both bully identification and classification case, the Random Forest algorithm with TFIDF embedding has performed better with an F1 score of 80.8 and 58.4 respectively.**

***Keywords*—Cyber bullying, Social Media, Machine Learning,**

I. Introduction

The advent of the internet and information technologies have revolutionized our way of communication, social life management, thinking, reasoning, etc. Especially in the past few years, online communication has migrated towards interactive social networking sites, blogs, mobile chat applications, etc. transcending all regional and spatial limitations through the internet. According to the Datareportal Global Overview report published on 27th January 2021, there are now 4.20 billion Social Media (SM) users around the world, which has grown by 490 million over the past 12 months [1]. The SM characteristics, such as accessibility, flexibility, being free, and having well-connected social networks, provide users with liberty and flexibility to post and write on their platforms has resulted in not only positive exchange of ideas but has also led to widespread dissemination of aggressive and potentially harmful content and has given these incidents an unprecedented power and influence to affect the lives of billions of people. Cyberbullying can be considered as a distinct phenomenon or as a sub-form of bullying with electronic devices [2]. A 2018 research study found that a majority of teens (59%) experienced some form of cyberbullying [3]. Asian countries with the most cyberbullying done in china (70%), Singapore (58%), and India (53%) [1]. Many government bodies and non-profit-healthcare organizations have highlighted the harmful effects of cyberbullying on the victims that include: depression, anxiety, reduced self-worth, difficulty sleeping, eating disorders, etc. based on psychological surveys across different countries [4]. So research on cyberbully detection on SM sites for its efficient monitoring and prevention is a high necessity of recent times.

The task of cyberbullying detection can be broadly defined as the use of Machine Learning (ML) techniques to automatically classify text in messages on bullying content or infer characteristic features based on higher-order information, such as user features or social network attributes. Therefore, the objectives of this research are: a) to explore ML algorithms to detect as well as classify cyberbullying from social media data, b) to propose ML based framework to prevent cyberbullying from Social Media. The paper has been structured as follows. The next section presents a brief overview of related works in the area of cyberbully detection. Section III details the conducted research methodology of the Twitter data for analysis purposes. Section IV provides an in-depth analysis and discussion of our proposed conceptual prevention framework. Section V provides a discussion on the findings and results of cyberbully detection by the adopted methods in this paper. The final section concludes the paper with discussions and limitations.

II. Literature Review

This section briefly discusses the past studies on cyberbullying detection. Hee et al. [5] conducted a study on automatic detection of cyberbullying in SM text where linear support vector machines (SVM) were explored. The classifier yielded an F1 score of 64% and 61% for English and Dutch cyberbully-related posts, respectively. Additionally, a study was done by Saravanaraj et al. [6] on automatic detection of cyberbullying from Twitter where Naïve Bayes and Random Forest classifiers were used to detect cyberbully-related posts and rumors were detected by using type and topic-specific classification and

978-1-6654-0637-6/21/$31.00 ©2021 IEEE

Twitter speech-act classifier. Moreover, Mihaylov and Nakov [7] studied hunting for troll comments in news forums.

Weller and Woo [8] did a study on Identifying Russian trolls on Reddit with deep learning and BERT word embedding. The study proposes a three-layer neural network architecture; their best model contains a Recurrent Convolutional Neural Network (RCNN) that outperforms current ML practices with an AUC of 84.6%. A study by Anzovino, Fersini, and Rosso in [9] automatic identification and classification of misogynistic language on Twitter proposed a taxonomy for modeling the misogyny phenomenon in online social environments with a Twitter dataset manually labeled.

Some studies have also been carried out to prevent and eliminate cyberbullying by applying the proposed detection method of bullying. A study was done by Sugandhi et al. [10] on automatic prevention and monitoring of cyberbullying designed a response grading system that maps an appropriate response for each cyberbully class taking into account the various parameters like the present social and political scenario, the severity, the overall sentiment against a particular issue, etc. A study conducted by Prime and Suri [11] on cyberbullying detection and prevention from data mining and psychological perspective stated the implementation of linear SVM (Support Vector Machine) with word vector and stemming process for sentiment analysis and bully measurement along with prevention measures such alerting SM site admin.

Much of the recent research uses small, heterogeneous datasets, without a thorough evaluation of applicability. Most cyberbully detection models can only detect cyberbullying; they cannot classify the type of bully. The reasons behind this could be the annotation scheme of data which is not trivial and that's why more fine-grained categories like insult, hatred, encouragement are sometimes hard to recognize. The algorithms concerning detection are mostly linear SVM classifiers or linear regression-based. Although huge research work has been on detection, no concrete prevention measure is mentioned in any of it. The little approach stated is not generalized to other platforms and languages beyond their collected dataset. Thus this study focuses on the framework for prevention of cyberbullying from SM by automatic detection of bully through exploration ML algorithms on the collected dataset from Twitter. Another concern of this research is to show a comparison of cyberbully detection by sentiment analysis (Bully identification model) and bully characteristics (Bully classification model) using Natural Language Processing (NLP) techniques.

III. Methodology

A. Developing ML Models

The ML models were built in the following stages: Data Collection, Data Preprocessing, Feature Extraction, Developing the Prediction Models, and Evaluating the Prediction Models. The phases are briefly discussed in the following subsections:

1) ***Data acquisition****:* The study is done only on English tweets. The public tweets related to cyberbullying are collected using publicly available application programming interface (API) worldwide. The collected cyberbully-related tweets were posted on Twitter between January 1, 2020, and May 31, 2021. This study used a total of 1000 labeled public tweets collected from the internet. Some of the keywords used here are "Isreal-Palestine", "feminism", "BLM", "rape", "pride month", "white-feminism", "racist" words, and most frequently used abusive languages.

2) ***Data Pre-processing****:* The acquired data is preprocessed through the cleaning of text data by removing username, URL, emojis, etc. using a python library tweet-preprocessor, filtering non-English words and HTML markups, the conversion of tweets to lowercase characters, and by removing tabs, and spaces. The stop words like "a", "an", "the" are removed using python's gensim library as they carry little meaning in a sentence. The stemming and tokenization have also been done to get them to a normalized state.

3) ***Data Annotation****:* From the collected dataset the corpora have been annotated for cyberbully detection and classification. For the detection of the cyberbully, the corpora have been annotated using the Valance Aware Dictionary and Sentiment Reasoner (Vader). lexicon [12]. For effective detection of a bully from text data a bully classification model is proposed which would be able to identify a bully according to its bullying characteristics. Hence a taxonomy has been designed which describes some specific textual categories related to cyberbullying, which include racial, offensive, misogynistic, and attacking comments targeted against the victims. All of these forms were inspired by social studies on cyberbullying and manual inspection of cyberbullying examples.

For the bully classification model, two anonymous annotators have annotated the same corpora independently according to the designed taxonomy for data validation. However, to measure how well the annotators made the same annotation decision for a certain category inter-annotator agreement has been measured using cohen's kappa score for pairwise annotator. The computed kappa score is around 0.83 which states the agreement between the annotators is almost perfect.

4) ***Feature Extraction****:* For this study embeddings like Term Frequency Inverse Document Frequency (TFIDF), Bag of Words (BoW) has been used for extracting feature vectors from the preprocessed tweets. Both embeddings are much simpler have been proved to perform better in machine learning algorithms.

5) ***Developing the Prediction Models****:* The bully identification model shall be able to classify text in two classes: positive sentiment and negative sentiment. The classification model classifies tweets in six classes: racism, offensive, attacking, misogyny, neutral, positive. To find the best detector model of bullying text different ML algorithms were chosen based on recent work relating to the detection of cyberbullying. In both cases, a random train test split of 70-30 was done. The models were developed using python's Scikit-learn library. Models of conventional ML algorithms such as Logistic Regression,

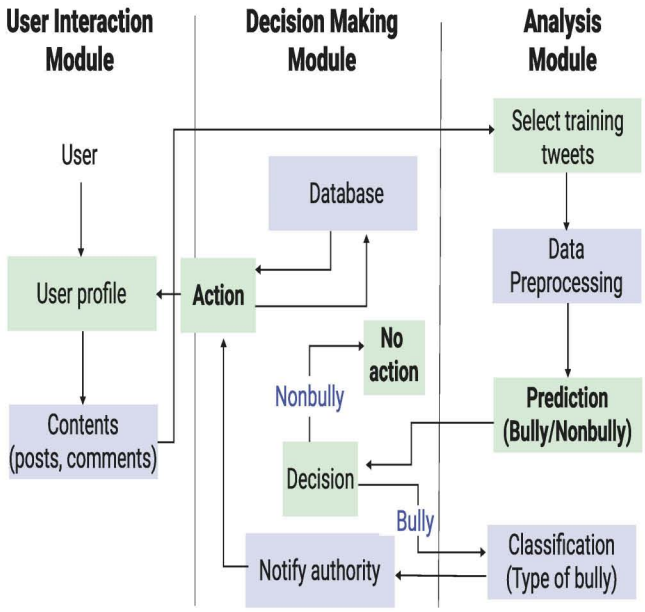

Fig. 1: Conceptual Framework

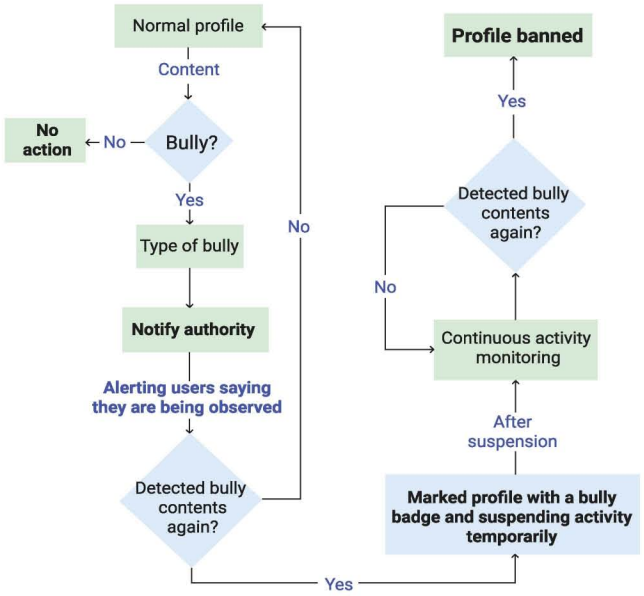

Fig. 2: Prototype of proposed framework

Multinomial Naïve Bayes, Random Forest, Linear SVM have been built on the vectorized dataset using TFIDF, BoW embedding techniques.

*6) **Evaluating the Models** :* The prediction models are generated using the training dataset whereas the performance of prediction models is evaluated for the unknown dataset (test set). The performance of bully identification models and bully classification models for the training and testing instances was measured in terms of precision, recall, and f1 score, and the results are shown in table I and table II respectively.

It is evident from table I and table II that the models with TFIDF embedding work better than BoW embeddings. In the BoW embedding technique Logistic Regression performed better whereas in TFIDF embedding Random Forest showed better accuracy for bully identification models. Again, for bully classification models Random Forest performed better for both embedding techniques.

IV. Conceptual Framework

To apply the cyberbully detection model, a conceptual framework is proposed and shown in Figure 1 and the prototype of this framework is shown in Figure 2.

There are three modules in this system which are consisted of the User interaction module, the Decision-making module, and the Analysis module. This proposal is essentially proposed to the authority of a social media website to abate cyberbullying in social media. After content is identified as a bully, the type of bully will be determined by the model. The model will give probabilities of the kinds of bullying mentioned in the taxonomy development section and the kind with the highest probability will be selected. The classification of the bully can be used according to the preference of a website, as different websites are built on different topics where the severity of the kinds of bully differs. When content is marked as a bully, the history will be collected against the user profile activity history (Database). If user posts or comments on a particular number of bullies within 24 hours, that user will be given an alert from the website letting them know that they are being monitored and authority will be notified with the victim profiles in a database. if the user bullies again in the next 168 hours user will be unable to post or comment for 168 hours and the user's profile will contain a badge showing that the user is a bully. The time allotment can be determined by authorities. If the user does not comment within a particular time, the suspension will be removed but the user will be marked as a bully in the user's history (Database). The bully badge will play a psychological role so that the user will not be willing to post bully content because that way they will gain the bully badge which is not a positive thing. After the action is over, if the user still comments on a bully, the profile will be banned permanently.

V. Discussion

Over the last decade, with new technological advancements, cyberbullying has become one of the most significant issues in the world. This study contributed to introducing a framework for preventing cyberbullying on SM by detecting bullying textual posts employing sentiment analyses and bullying features through the exploration of ML algorithms. The dataset corpora extracted from Twitter are diverse in content. The taxonomy designed for bully characteristics covers most of the areas a person is bullied for such as racism, prejudice against women, disrespectful or insulting words, aggressiveness, etc. Embedding methods such as BoW and TFIDF have been

TABLE I: Performance of Bully Identification models in terms of precision,recall and F1 scores

Model		Train Set			Test Set		
Embedding	**Algorithm**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
BOW	Log regression	0.832	0.830	0.830	0.782	0.779	0.778
	Naive Bayes	0.765	0.765	0.765	0.723	0.723	0.723
	SVM	0.837	0.834	0.834	0.774	0.769	0.768
	RF	0.893	0.883	0.882	0.750	0.747	0.746
TFIDF	Log regression	0.892	0.892	0.892	0.794	0.794	0.794
	Naive Bayes	0.873	0.873	0.873	0.742	0.742	0.742
	SVM	0.910	0.909	0.909	0.781	0.781	0.781
	RF	0.943	0.937	0.937	0.813	0.809	0.808

TABLE II: Performance of Bully Classification models in terms of precision,recall and F1 scores

Model		Train Set			Test Set		
Embedding	**Algorithm**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BOW	Log Regression	0.823	0.759	0.771	0.617	0.515	0.515
	Naive Bayes	0.766	0.705	0.714	0.571	0.491	0.490
	SVM	0.860	0.792	0.805	0.648	0.506	0.511
	RF	0.872	0.806	0.818	0.637	0.527	0.524
TFIDF	Log Regression	0.834	0.838	0.835	0.626	0.559	0.576
	Naive Bayes	0.835	0.827	0.815	0.553	0.543	0.544
	SVM	0.888	0.885	0.886	0.596	0.489	0.505
	RF	0.980	0.979	0.979	0.635	0.575	0.584

used on thoroughly preprocessed tweets for applying selected classifiers namely: Liblinear based Logistic Regression, Linear SVM, Multinomial Naïve Bayes, and Random Forest. It is evident from several performance evaluation measures of the models that the Random Forest model with TFIDF embedding performs better in both cases. The highest achieved accuracy (F1 score) for the bully identification model is 80.8% and for the bully classification model is 58.4%. The framework proposed for the prevention of cyberbullying is a prevention measure generalized for all SM platforms. In the proposed framework, the analytics module shall automatically analyze user posts/comments through deployed ML models in the cloud and the decision module shall take action accordingly if a bullying attempt is found. This framework is targeted to contribute for SM authorities so that monitoring of bullies can be automatic and certain measures can be taken against the guilty to stop cyberbullying trend for good.

However, this paper used only two NLP techniques named BoW and TFIDF to transform text data into machine-readable vectors for feature extraction. Again, for both identification and classification of bullies, a limited amount of 1000 labeled tweets has been used exploring only the ML algorithms. The use of modern word embedding techniques along with neural networks could have improved the performance of the models. Furthermore, dataset corpora have been used is collected from Twitter only including the publicly available tweets as it provides a public API for data extraction. Other SM platforms such as Facebook could have been investigated as it is the most used communication media worldwide with a record of severe cyberbullying. For future works, we aim to work on these limitations. Additionally, we would also like to contribute to cyberbully detection from trolls and memes so that we can have harmless and safer SM.

REFERENCES

[1] " DIGITAL 2021: GLOBAL OVERVIEW REPORT," availableon:https://datareportal.com/reports/digital-2021-global-overview-report, last accessed: 12 July, 2021.

[2] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *Journal of child psychology and psychiatry*, vol. 49, no. 4, pp. 376–385, 2008.

[3] M. Anderson, "A majority of teens have experienced some form of cyberbullying," 2018.

[4] " EFFECTS OF CYBERBULLYING," availableon:https://americanspcc.org/impact-of-cyberbullying/, last accessed: 22 July, 2021.

[5] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PloS one*, vol. 13, no. 10, p. e0203794, 2018.

[6] A. Saravanaraj, J. Sheeba, and S. P. Devaneyan, "Automatic detection of cyberbullying from twitter," *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2016.

[7] T. Mihaylov and P. Nakov, "Hunting for troll comments in news community forums," *arXiv preprint arXiv:1911.08113*, 2019.

[8] H. Weller and J. Woo, "Identifying russian trolls on reddit with deep learning and bert word embeddings," 2019.

[9] M. Anzovino, E. Fersini, and P. Rosso, "Automatic identification and classification of misogynistic language on twitter," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2018, pp. 57–64.

[10] R. Sugandhi, A. Pande, A. Agrawal, and H. Bhagat, "Automatic monitoring and prevention of cyberbullying," *International Journal of Computer Applications*, vol. 8, pp. 17–19, 2016.

[11] S. Parime and V. Suri, "Cyberbullying detection and prevention: Data mining and psychological perspective," in *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]*. IEEE, 2014, pp. 1541–1547.

[12] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014.