AUTOMATIC DETECTION OF CYBERBULLYING FROM TWITTER

A.Saravanaraj
PG Student
Department of CSE
Pondicherry Engineering College, Puducherry-605014
9042622962
saravanaraj@pec.edu

J. I. SheebaAssistant ProfessorDepartment of CSE
Pondicherry Engineering College, Puducherry-6050149443084976sheeba@pec.edu

S. Pradeep Devaneyan
Dean, School of Mechanical and Building Sciences
Christ College of Engineering and Technology
Puducherry-605010
9047007444
pr.signs@gmail.com

***Abstract*—With the increasing use of social communication networks by their users leads to huge amount of user-generated communication data. Due to the popularity of social media cyberbullying become the major problem in online communication and cyberbullying behavior received more and more attention. Cyberbullying may cause many serious and negative impacts on person's life and even leads to teen suicide. In the existing system the set of unique features derived from Twitter such as network, activity, user and tweet contents. By using these features the cyberbullying words which are presented in the tweet contents are detected using machine learning algorithms. The rumor tweets are detected using syntactic and semantic techniques. The cyberbully detection and rumor detection on twitter network are done separately in the existing technique. In the proposed work the detection of cyberbully words and rumor tweets on twitter are integrated into a single application, along with these the cyberbully contents in the tweet comments will be detected using Naïve Bayes and Random Forest classifier. The name, gender and age of the cyberbully tweeted people will also be detected using feature extraction technique. By using type and topic specific classification and Twitter speech-act classifier, the rumor tweets will also be detected in this proposed work.**

Keywords-Cyberbullying Detection; Machine learning algorithms; Twitter; Feature extraction; Rumor detection.

I. INTRODUCTION

Social networking sites have become immensely popular in the last few years. More than millions of users have used these websites as communication tools and as real-time, dynamic data sources. Where users can create their own profiles and communicate with other users regardless of location and physical limitations. Cyber criminals have utilized social media as a new platform in committing different types of cybercrimes such as phishing, spamming, spread of rumors, and cyberbullying**.**

In particular, cyberbullying and rumors has emerged as a major problem along with the recent development of online communication and social media. Cyberbullying can be defined as the use of information and communication technology by an individual or a group of users to harass other users. Cyberbullying has also been extensively recognized as a serious national health problem, in which victims demonstrate a significantly high risk of suicidal ideation. Cyberbullying is a substantially persistent version of traditional forms of bullying with negative effects on the victim.Social media provides users not only a good platform for communication and information sharing, but also an easy access to fresh news. However,these platforms are also places where users experience bullying as victims, bullies or bystanders. One study conducted by national anti-bullying charity has shown that two out of three 13-22 years old who were surveyed have been victims of cyberbullying [1, 2]. Applying machine learning may provide successful or unsuccessful cyberbullying predication results, because building a successful machine learning model depends on many factors and the features extracted from the twitter network are used to train a machine learning algorithm for getting the effective results. The Naïve Bayes and random forest classifiers is used to detect the cyberbullying content which are present in the tweets and comments.

Rumors is also considered as a serious problem in social networks along with the cyberbullying. There are widely varying definitions of the term "rumor", a rumor could be both true and false. A rumor is a claim whose truthfulness is in doubt and has no clear source, even if its ideological or partisan origins and intents are clear [4]. Due to the development of social network, the amount of information has been growing explosively. However, the quality of information does not become better. All kinds of false information, especially rumors, have permeated almost every corner of social networking sites. Therefore, automatic assessment of information credibility has received considerable attention in last years. Detection of rumors is one of the critical research topics of information credibility. It is often viewed as a tall tale of explanations of event circulating from person to person and pertaining to an object, event, or issue in public concern. The rumor that spreads in the twitter networks create a severe issues on the victims side so along with the cyberbully detection on the twitter network the detection of rumor is also important but finding the rumors is a complex task [3]. In this proposed work, different types of aspects will be considered to overcome the different issues on detecting rumors. The type and topic specific classification and Twitter speech-act technique, Naïve Bayes classifier will be used to detect the rumors on twitter. In this paper section II describes the related works, section III describes the proposed work and section IV describes the conclusion of the paper.

II. RELATED WORKS

This section contains brief reviews about some existing works in cyberbullying and rumor detection.

Chen et al. developed an approach for bullying detection that was equipped with a lexical syntactic feature, Although lexical features perform well in detecting offensive entities without considering the syntactical structure of the whole sentence, they fail to distinguish sentence offensiveness which contain same words but in different orders [2]. Using data sets from Myspace, Dadvar et al. developed a gender based bullying detection approach that used the gender feature in enhancing the discriminative capacity of a classifier, not all the users provide complete information it leads to the imbalancing of the datasets it affects the efficiency of the model [6]. Nalini and Sheela proposed an approach for detecting cyberbullying messages in Twitter by applying a feature selection weighting scheme and latent dirichlet allocation [7]. Chavan and Shylaja included pronouns, skip-gram, TF-IDF, and N-grams as additional features in improving the overall classification accuracy of their model, the TF-IDF is not semantic [8].

There are a large number of related studies on rumor detection. Most works focused on detecting rumors by shallow features of messages, including content, blog features. This method obtains significant improvement, compared with the state-of-the-art approaches. But such shallow features cannot differentiate the rumor messages from normal messages in many cases [3]. Sardar Hamidian and Mona Diab illustrates rumor detection and classification, natural language processing tools are used for detection, classification and verifying, there are four major aspects to be checked: Provenance, Source, Date and location. Verifying the trustworthiness of the data is complex [4]. Another detection model use syntactic and semantic feature which gives an effective result, the accuracy is above 90%. An extensive literature review has been made on above existing techniques Based on the above an accurate method is needed for predicting the above tasks. Hence to overcome the limitations of existing techniques, intelligent text mining techniques will be proposed by incorporating rumor detection and cyberbully detection from the twitter with the aim of providing accurate results and less error rate.

III. PROPOSED WORK

The Figure 1 represents framework for the proposed work which can be explained as follows: the process of detecting cyberbully words and the rumor tweets from input dataset. The input twitter dataset is collected from the given (web address) link http://lsir.epfl.ch/research/datasets. An Input dataset is sent to data preprocessing which is applied to improve the quality of the input data. The data preprocessing also includes removing stop words and special characters. After performed the data preprocessing the output data is sent to classification algorithm for detecting the cyberbully words in tweets, tweet comments and also it will find the rumor tweets. Finally the name, gender and age of the cyberbully tweeted people are extracted from the input data. The proposed work will detect Cyberbully words on tweets, comments and Rumor from input dataset, which includes 3 Steps,

A. Data Preprocessing

B. Cyberbully Detection

C. Rumor Detection

A. Data Preprocessing

Social network data are noisy, thus preprocessing has been applied to improve the accuracy of the input data. This includes removing stop words. Stop words are usually like “a”, “as”, “have”, “is”, “the”, “or”, etc. Stop words mainly used for consumed memory space and processing time.

B. Cyberbully Detection

In cyberbully detection, the bullying words in the tweet contents and tweet comments are detected using the machine learning algorithms. After getting the output from the preprocessing step, the output file will sent to the classification algorithms. There the trained classifier will be used for detection. The training dataset consist of list of cyberbullying words. With the training dataset the preprocessed twitter dataset is tested for bullying word presence or not. The Naïve Bayes and the Random Forest classifiers are mainly used to detect the cyberbully words present in the tweet contents and comments. This method will also be identify the information of cyberbully tweeted people.

1. *Naïve Bayes*

The Naive Bayes classifier technique is based on Bayesian theorem with independence assumptions between predictors. The classifier is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for huge datasets.Assume that the presence of absence of a particular feature of a class is unrelated to the presence or the absence of any other feature. Each word in a tweet is considered as a unique variable in the Naive Bayes classifier to determine the probability of that word and whether it belongs to class: present or absent [6, 13]. The classifierscan be greatly simplified by assuming that features asindependent class, that is

$$P(c \mid x) = P(x \mid c) * P(c) / P(x)$$

Where, variable P is a probability, variable c is a class and variable $x=\{x_1, x_2 \ldots x_n\}$ is a vector.

2. *Random Forest*

Random forest classifier is an ensemble learning technique for classification and regression, which operate by constructing a lot of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

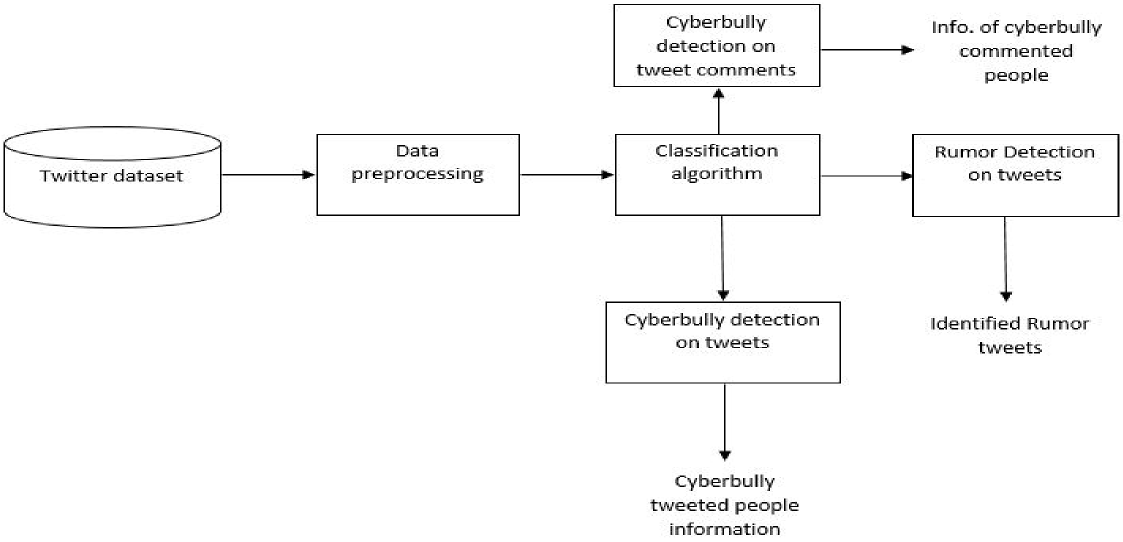
Twitter dataset
Data preprocessing
Classification algorithm
Cyberbully detection on tweet comments
Info. of cyberbully commented people
Rumor Detection on tweets
Identified Rumor tweets
Cyberbully detection on tweets
Cyberbully tweeted people information

Figure 1. Architecture Diagram for Proposed Work

C. *Rumor Detection*

In rumor detection, the detection of rumors will be taken as a classification problem and detected by using classification techniques such as Naïve Bayes, type and topic specific classification technique and Twitter speech-act classifier. Finally, the rumor tweeted people information are extracted from the input dataset.

1. *Twitter speech-act classifier*

In twitter speech act classifier, there are six speech act categories that are commonly seen on twitter. They are Assertion: An assertion is a tweet that commits the speaker to the truth of the expressed proposition.So an assertion can be assessed by the truth value of its proposition;Recommendation: Tweets that recommend things (such as links) or give advice about a situation fall under this category; Expression: Tweets that express the speaker's attitudes and emotions towards something; Question: Tweets that are asking for information or confirmation lie under this category; Request: Requests are tweets that attempt to get the hearer to do or stop doing something; Miscellaneous: All remaining speech act types grouped into this miscellaneous category [11,12].

2. *Type and topic specific classification*

The definitions for topic and type, a topic is a subject discussed in one or more tweets (e.g.: Boston Marathon bombings, Red Sox, global warming, etc.). The type characterizes the nature of the topic. The identified three topic types on twitter such as Entity-oriented topics: Topics about entities such as celebrities, brand names, sportsteams, etc.; Event-oriented topics: Topics about events in the world, most commonly about breaking news; Long-standing topics: topics about subjects that are commonly discussed in everyday talk, such as music, global warming, cooking, travelling, etc. [12].

By using above techniques, the machine learning algorithms will be trained to detect the cyberbully words and rumor in twitter network.

IV. CONCLUSION

The proposed work will be focused on detecting the occurrence of cyberbullying and rumor in twitter networks using machine learning algorithms, type & topic specific classification and Twitter speech-act classifier, finally the cyberbully tweeted and rumor spreading people information will also be extracted. And the integration of cyberbully detection and rumor detection in single application makes the detection easier. This proposed model may give better result in preventing the users of the social networkfrom becoming victims when compared other existing techniques.

REFERENCES

[1] Rui Zhao, Anna Zhou, Kezhi Mao, Automatic Detection of Cyberbullying on Social Networks based on Bullying Features, ICDCN '16 Article No. 43, January 2016, ACM.

[2] Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu. "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), pp. 71-80. IEEE, 2012.

[3] Qiao Zhang, Shuiyuan Zhang, Jian Dong, Jinhua Xiong, and Xueqi Cheng.Automatic Detection of Rumor on Social Network,pp. 113–122, Springer (2015).

[4] Sardar Hamidian and Mona Diab. Rumor Detection and Classification for Twitter Data, IARIA (2015), 71-77, SOTICS 2015: The Fifth International Conference on Social Media Technologies, Communication, and Informatics, ISBN: 978-1-61208-443-5.

[5] Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In Advances in information retrieval (pp. 693-696). Springer.

[6] Mohammed Ali Al-garadi, Kasturi DewiVarathan, Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network, Computers in Human Behavior 63 (2016) 433-443, Elsevier.

[7] Nalini, K., & Sheela, L. J. (2015). Classification of Tweets using text classifier to detect cyber bullying. In Emerging ICT for bridging the future-Proceedings of the 49th Annual convention of the Computer Society of India CSI (Vol. 2, pp. 637-645). Springer.

[8] Chavan, V. S., & Shylaja, S. (2015). 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 (pp. 2354-2358). IEEE.

[9] Nafsika Antoniadou, Constantinos M. Kokkinos, Angelos Markos. Possible common correlates between bullying and cyber-bullying among adolescents, Psicología Educativa 22 (2016) 27–38, Elsevier.

[10] P.V.Bindu, P.Santhi Thilagam. Mining social networks for anomalies: Methods and challenges, Journal of Network and Computer Applications 68(2016)213–229, Elsevier.

[11] X. Zhao and J. Jiang. An empirical comparison of topics in twitter and traditional media. Singapore Management University School of Information Systems Technical paper series. Retrieved November, 10:2011, 2011.

[12] Vosoughi, Soroush, and Deb Roy. Tweet acts: A speech act classifier for twitter. arXiv preprint arXiv:1605.05156 (2016).

[13] Sanchez, Huascar, and Shreyas Kumar. "Twitter bullying detection." ser. NSDI 12 (2011): 15-15.