AGGRANDIZED TECHNIQUE FOR RECOGNISING OF OFFENSIVE LANGUAGE IN ONLINE NETWORK

K. Shantha Kumari, R. Prem Kumar, Mohammed Thowfeeq, Prajeesh Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology Puducherry, India. shanthajayakumar@gmail.com
K. Mahalakshmi, R Keerthika, Information Technology, Karpagam College of Engineering, Coimbatore, India mahalakshmi.k@kce.ac.in

ABSTRACT

Extensive usage of Social media in this modern era provides entirely a new form of social interaction and communication especially among the adolescent and youth population. They tend to converge at social media for sharing their opinions and thoughts on common issues. However, the young people are also vulnerable to cyber-bullying by means of the offensive language which is spilled across the online web. To avoid severe results (mental trauma, stress etc.), it is important to act proactively to detect online bullying activity. This paper discusses an optimized technique to recognize the offensive language in Online Social media, specifically for Twitter. Previous works mostly concentrated on Keyword matching along with intelligent technologies to recognize the whether the published content is cyberbullied or not. In addition, most of the algorithms were tried only on stored dataset. Cyberbullying can be thwarted if the probability of a tweet user to be an abuser could be predicted earlier with a live capturing of data’s from Twitter. This research work proposes an optimized technique for detecting the offensive language in Tweets by Naïve Bayesian Text Classification. In addition, FGA logic is used to detect the possibility of abusive Tweet users. This approach is evaluated with data’s from Twitter and the performance of this optimized technique is discussed.

Index Terms— Abuse, Classification, Cyberbullying, Twitter

I. INTRODUCTION

Social media use web-based and mobile technologies on smart phones and tablet computers to create highly interactive platforms through which individuals, communities and organizations can share, co-create, discuss and modify user-generated content or pre-made content posted online [1]. Cyberbullying is an activity which is been played through technology to send mean, threatening, or embarrassing messages to or about another person. It might be in a text, message, or in a post through social media. It also has a wider audience, and can spread quickly. There are more number of victims who are affected due to this activity and the consequences for cyberbullying might be fatal too. Cyber bullying activities includes Harassment, Racism, Threat, Character assignation etc. in Online Social Media. So, the monitoring and controlling works regarding the online bullying activity has been increasing rapidly around social media. At any cost, Cyber bullying need to be prevented. Hence the abuse posts need to be detected in order to protect cyberbullying activity. The Existing methodologies that detect cyberbullying use several techniques: SVM, LSF, MCES, Text mining and etc.

The above-mentioned works are able to distinguish good posts by Keyword matching based on stored social media datasets. However, cyberbullying detection would be efficient if the “live post” could be classified as it is being streamed. Moreover, there is a need to find out the probability for a person to be cyberbullying activist based upon his/her previous posts.

In this work, an optimized technique using Naïve Bayesian classifier and FGA logic is proposed. Along with these intelligent techniques, Twitter API is used to stream the live feed and analyzed using Dictionary method. To identify the cyberbullying criminal, LIWC method is adapted. By this method, overall words used by a person is monitored and continuously compared to identify what he/she is typing in the social media.

II. RELATED WORK

There are a lot of literatures related to event detection and cyberbullying detection system been reviewed and studied to gain insight of the system development and effectiveness. The related works talks about the Cyberbullying detection and prevention in the social media in which the persons who were involved in bullying activities.

The authors in [2] used Applied rule-based learning to develop a model for detecting cyberbullying based on textual features (e.g., the number of curse words in a message) and compared its performance to a bag-of-words model (i.e., based on a matrix of all the words that occur in the training corpus). The work explained in [3] speaks about the specific choice of words and subtle structure of sentences can persuade the reader towards one point of view or another and are sufficient to influence whether people interpret violent acts as patriotism or terrorism. Lexi-con approaches utilize wordlists containing known profane words to match them against a given text. In their naïve form, they classify a text as online harassment, if it contains offending words. The classification performance varies considerably depending on the wordlist used [4]. Another work takes Lexical Syntactic Feature (LSF) architecture for people involved in the act like using offensive language in social media [5]. Abusive users who are ready to bully another person via social media. The proposed system is explained as follows: the work explained in [6] speaks about the expert knowledge for automatic detection of bullies by the person involved in harassment in use of Multi-
Criteria Evaluation System (MCES). The researchers in [7] applied a hybrid approach combining supervised models with an expert system work predominantly employs lexicon and machine learning approaches given in. The authors in [8] assert that there are many techniques which are useful for text document classification and point out the support vector machine (SVM) algorithm as superior. The methodology in [9] described bullying detection using intelligence techniques work to detect the cyberbullying victims by the FuzGen learning algorithm. In addition, the work in [10] details a system which consists of modules such as Link filtering, age validation and comments validation by which unwanted comments are blocked. The work in [11] details a model that simultaneously discovers instigators and victims of bullying as well as new bullying vocabulary by starting with a corpus of social interactions and a seed dictionary of bullying indicators. These are the major related works which speaks about the Cyberbullying detection. We examined these work to produces an optimized detection process with the respective techniques.  
III. PROPOSED SYSTEM

The proposed system focuses on two important objectives:

1. To capture live tweets and compare them with stored dictionary and detect the abusive tweets  
   The live tweets are tokenized and compared with user-build dictionary. The tweets are generally may follow formal and informal language. By tokenizing the sentence is broke down into individual words. These words are now compared and classified using Naive Bayesian classifier. The advantage of this classifier is that it provides most accurate results.

2. To predict the whether the user could be a cyber-bullying person based on their post [type of words]

Any person who is prone or accustomed to use bad words repeatedly in real life has the habit to reflect the same in Twitter. Building upon this assumption, the tweets [tokens] after classification are counted. When the abusive words crossed certain limits, then the user can be classified as above. Abusive users who are ready to bully another person via social media.  
IV. METHODOLOGY

This model implementation is based on the python with the SQL as the backend. Twitter API and Rapid Miner classifier are used. This system detects the cyberbullying related tweets which is matched with the keywords in the database.

![Flow Diagram](image)

**Fig 2: Flow Diagram**

The different phases of the proposed model is explained as follows:  
**Capturing Live tweets:**  
Twitter Stream APIs are available that open the door for capturing live feed from Twitter. During any sensitive issue, tweets will be flooded, which can be streamed using Public Stream API defined by Twitter Development Documentation [13] as shown in Figure 3:

![Public Streaming API working](image)

**Fig 3: Public Streaming API working**

The captured live tweets are stored in SQL Server for further processing.  
**Naive Bayesian Classification [12]:**  
A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes’ theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be independent feature
model. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

The Naive Bayes Classifier is based on the bag-of-words model. With the bag-of-words model the words of the text-document are checked to appear in a positive-words-list or a negative-words-list. If the word appears in a positive-words-list the total score of the text is updated with +1 and vice versa. If at the end the total score is positive, the text is classified as positive and if it is negative, the text is classified as negative.

**Fuzzy C-Means [13]:**
Fuzzy c-means (FCM) is a data clustering technique in which a dataset is grouped into c clusters with every data point in the dataset belonging to every cluster to a certain degree.

**add cluster attribute**
If true, the cluster id is stored in an attribute with the special role 'label' instead of 'cluster'.

**Type: Boolean Default: true add as label**
If true, the cluster id is stored in an attribute with the special role 'label' instead of 'cluster'.

**Type: Boolean Default: false Add partition matrix**
If true, results also contain membership matrix which shows how strong given example belong to particular cluster.

**Type: boolean Default: false**
This system has the process of detection of abusive word and checks whether the user is cyber bully person or not by getting input from the social Network. The tweets are used as input, where the process starts of the proposed system and the data's are streamed with twitter API. By utilizing the Twitter APIs, the cyberbullying related tweets will be retrieved by the connection of the APIs. The data's which are streamed through twitter API are processed with an algorithm to check whether the tweets are good or bad.

While streaming the tweets are processed by tokenizing method, the tokenized tweets are matched based on the cyberbullying keywords in the database, the tweets of the respective user are predicted whether the user is bullied person are not, with the help of LIWC (Linguistic Inquiry and Word Count) and bad tweets are detected. The detected cyberbullying related tweets are stored in the database and it’s fed into data mining process such as classifying, clustering and followed by fuzzy rule set. The data’s after this process are been normalized and stored in the database. The proposed system process tells how the cyberbullying related tweets is detected and reported to the concern.

### V. EVALUATION

To evaluate the proposed model tweets are extracted from twitter. The bad content tweets are stored in the database. The filtered tweets are then used for classification. The tweets are classified into categories:

**Category - Description**

**Good Tweet** - The algorithm checks whether the tweet is Good. **Bad Tweet** - The algorithm checks whether the tweet is Bad.

**Bullied User** - The users involved in cyberbullying activity Frequently. **Normal User** - The users not involved in bullying activity.

The following figure visualizes the Different types of tweets as explained above using various colors:

![Fig 4: Abusive tweet](http://example.com/image.png)

The good tweets are eliminated, and this system focuses on the bad tweets which is stored in the database. The social media users who are involved in the cyberbullying acts are predicted that they are cyberbullied person or not, with the help of LIWC method. It’s focusing on the tweets posted by the users in Twitter and capture the keywords written by the users for keyword matching in order to determine the cyberbullying event, the identity of the cyberbullies and their details who are involved frequently. The following snapshot presents the LIWC output which aids in identifying the Abuser:

![Fig 5: Abuser](http://example.com/image.png)

The system only focuses on the text posted by the users in the social network, the emotional icons will not be taken into account. We are focusing on large
types of cyberbullying related word that will appear in our targeted tweets. The social network platform that we are going to further discuss and research on will be in the context of Twitter. We are focusing on the tweets posted by the users in Twitter and capture the keywords written by the users for keyword matching in order to determine the cyberbullying event, the identity of the cyberbullies and their details. Future work is that it moves along with the prevention of the cyberbullying related tweets and addition of emotional icons will be accounted in the detection process. As the result this system shows the cyberbullying users and related tweets in the social network and reported to the alerting agency of the cybercrime department.

VI. CONCLUSION
The system only focuses on the text posted by the users in the social network, the emotional icons will not be taken into account. We are focusing on large types of cyberbullying related word that will appear in our targeted tweets. The social network platform that we are going to further discuss and research on will be in the context of Twitter. We are focusing on the tweets posted by the users in Twitter and capture the keywords written by the users for keyword matching in order to determine the cyberbullying event, the identity of the cyberbullies and their details. Future work is that it moves along with the prevention of the cyberbullying related tweets and addition of emotional icons will be accounted in the detection process. As the result this system shows the cyberbullying users and related tweets in the social network and reported to the alerting agency of the cybercrime department.

VII. REFERENCES