AGGRANDIZEDTECHNIQUEFOR 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 Naive 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 table 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 activeity 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 controling 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 developpment 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 Naïve Bayesian classifier. The advantage of this classifier is that it provides most accurate results.

The proposed system is explained as follows

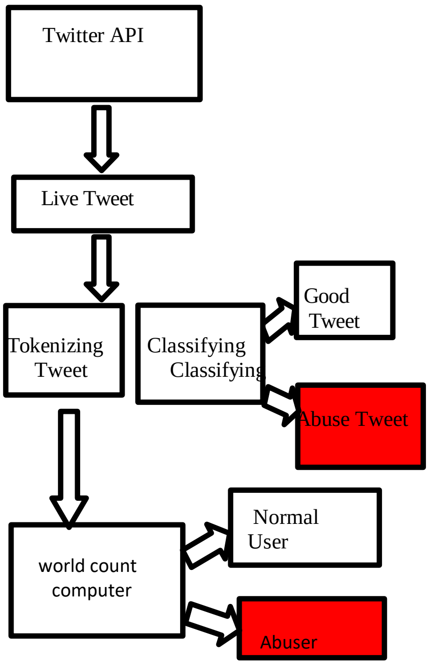


Fig 1: Proposed system for Abusive user prediction

2. To predict the whether the user could be a cyberbullying 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.

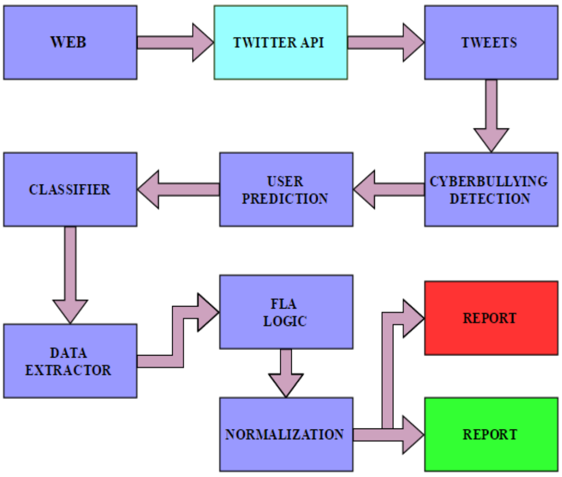


Fig 2: Flow Diagram

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:

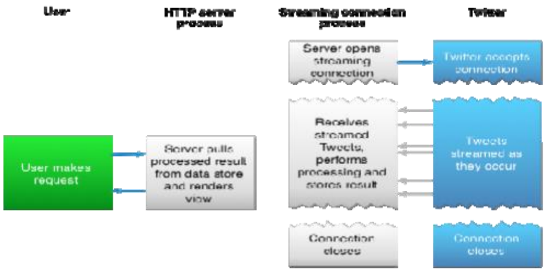

User
HTTP server process
Streaming connection process
Twitter
Server opens streaming connection
Twitter accepts connection
User makes request
Server pulls processed result from data store and renders view
Receives streamed Tweets, performs processing and stores result
Tweets streamed as they occur
Connection closes
Connection closes

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 n 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 cyber-bullying related tweets will be retrieved by the conection 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

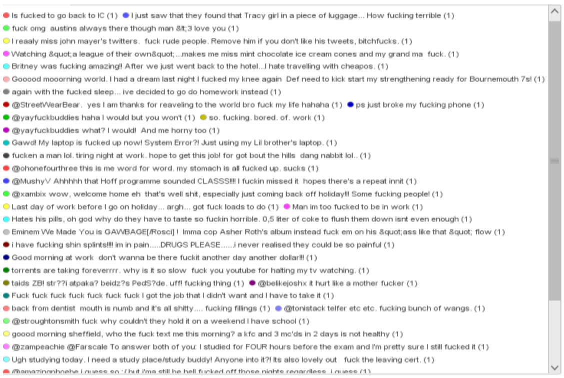
To evaluate the proposed model tweets are extracted from twitter. The bad content tweets are stored in the database. The filtered tweets are then used for classification. The tweets are classified into categories:

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:



● Is fucked to go back to IC (1) ● I just saw that they found that Tracy girl in a piece of luggage... How fucking terrible (1)

● fuck omg austins always there though man &lt;3 love you (1)

● I reaaly miss john mayer's twitters. fuck rude people. Remove him if you don't like his tweets, bitchfucks. (1)

● Watching &quot;a league of their own&quot;...makes me miss mint chocolate ice cream cones and my grand ma fuck. (1)

● Britney was fucking amazing!! After we just went back to the hotel...I hate travelling with cheapos. (1)

● Gooood mooorning world. I had a dream last night I fucked my knee again Def need to kick start my strengthening ready for Bournemouth 7s! (1)

● again with the fucked sleep... ive decided to go do homework instead (1)

● @StreetWearBear. yes I am thanks for reaveling to the world bro fuck my life hahaha (1) ● ps just broke my fucking phone (1)

● @yayfuckbuddies haha I would but you won't (1) ● so. fucking. bored. of. work (1)

● @yayfuckbuddies what? I would! And me horny too (1)

● Gawd! My laptop is fucked up now! System Error?! Just using my Lil brother's laptop. (1)

● fucken a man lol. tiring night at work. hope to get this job! for got bout the hills dang nabbit lol.. (1)

● @ohonefourthree this is me word for word. my stomach is all fucked up. sucks (1)

● @MushyV Ahhhhh that Hoff programme sounded CLASSS!!!! I fuckin missed it hopes there's a repeat innit (1)

● @xambix wow, welcome home eh that's well shit, especially just coming back off holiday!!! Some fucking people! (1)

● Last day of work before I go on holiday... argh... got fuck loads to do (1) ● Man im too fucked to be in work (1)

● Hates his pills, oh god why do they have to taste so fuckin horrible. 0,5 liter of coke to flush them down isnt even enough (1)

● Eminem We Made You is GAWBAGE[/Rosci] ! Imma cop Asher Roth's album instead fuck em on his &quot;ass like that &quot; flow (1)

● i have fucking shin splints!!!! im in pain.....DRUGS PLEASE......i never realised they could be so painful (1)

● Good morning at work don't wanna be there fuckit another day another dollar!!! (1)

● torrents are taking foreverrrr. why is it so slow fuck you youtube for halting my tv watching. (1)

● taids ZB! str??i atpaka? beidz?s PedS?de. uff! fucking thing (1) ● @belikejoshx it hurt like a mother fucker (1)

● Fuck fuck fuck fuck fuck fuck fuck I got the job that I didn't want and I have to take it (1)

● back from dentist mouth is numb and it's all shitty.... fucking fillings (1) ● @tonistack telfer etc etc. fucking bunch of wangs. (1)

● @stroughtonsmith fuck why couldn't they hold it on a weekend I have school (1)

● goood morning sheffield, who the fuck text me this morning? a kfc and 3 mc'ds in 2 days is not healthy (1)

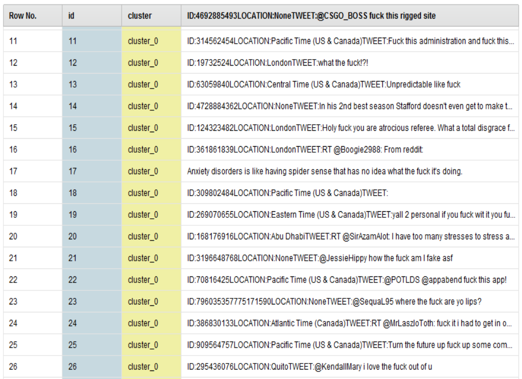
● @zampeachie @Farscale To answer both of you: I studied for FOUR hours before the exam and I'm pretty sure I still fucked it (1)

● Ugh studying today. I need a study place/study buddy! Anyone into it?! Its also lovely out fuck the leaving cert. (1)

● @amazingphoebe i guess so :/ but i'ma still be hell fucked off those nights regardless. i guess (1)

Fig 4: Abusive tweet

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 cyber-bullied 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 cyber-bullies and their details who are involved frequently. The following snapshot presents the LIWC output which aids in identifying the Abuser:



Row No.	id	cluster	ID:4692885493LOCATION:NoneTWEET:@CSGO_BOSS fuck this rigged site
11	11	cluster_0	ID:314562454LOCATION:Pacific Time (US & Canada)TWEET:Fuck this administration and fuck this...
12	12	cluster_0	ID:19732524LOCATION:LondonTWEET:what the fuck!?!
13	13	cluster_0	ID:63059840LOCATION:Central Time (US & Canada)TWEET:Unpredictable like fuck
14	14	cluster_0	ID:4728884362LOCATION:NoneTWEET:In his 2nd best season Stafford doesn't even get to make t...
15	15	cluster_0	ID:124323482LOCATION:LondonTWEET:Holy fuck you are atrocious referee. What a total disgrace f...
16	16	cluster_0	ID:361861839LOCATION:LondonTWEET:RT @Boogie2988: From reddit:
17	17	cluster_0	Anxiety disorders is like having spider sense that has no idea what the fuck it's doing.
18	18	cluster_0	ID:309802484LOCATION:Pacific Time (US & Canada)TWEET:
19	19	cluster_0	ID:269070655LOCATION:Eastern Time (US & Canada)TWEET:yall 2 personal if you fuck wit it you fu...
20	20	cluster_0	ID:168176916LOCATION:Abu DhabiTWEET:RT @SirAzamAlot: I have too many stresses to stress a...
21	21	cluster_0	ID:3196648768LOCATION:NoneTWEET:@JessieHippy how the fuck am I fake asf
22	22	cluster_0	ID:70816425LOCATION:Pacific Time (US & Canada)TWEET:@POTLDS @appabend fuck this app!
23	23	cluster_0	ID:796035357775171590LOCATION:NoneTWEET:@SequaL95 where the fuck are yo lips?
24	24	cluster_0	ID:386830133LOCATION:Atlantic Time (Canada)TWEET:RT @MrLaszloToth: fuck it i had to get in o...
25	25	cluster_0	ID:909564757LOCATION:Pacific Time (US & Canada)TWEET:Turn the future up fuck up some com...
26	26	cluster_0	ID:295436076LOCATION:QuitoTWEET:@KendallMary i love the fuck out of u

Fig 5: Abuser

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

[1] S. Sahu, S. K. Nanda and S. Baral, Social Media as A Tool for Marketing and Promotion of Library Information, NIT Rourkela, (2016).

[2] R.R.A.H.L.K. Dinakar, Modeling the Detection of Textual Cyberbullying, *5th International AAAI Conference on Weblogs and Social Media* (2011).

[3] E.M.M. &. N. Dunn, The War of the Words: How Linguistic differences in Reporting shape Perceptions of Terrorism. The society for the Psychological Study of Social (2012).

[4] S.O.C.E.F. &. A.J. Sood, Automatic identification of personal insults on social news sites. Journal of the American Society for Information Science and Technology 63(2): 270-285 (2012).

[5] Y. z. z. x. Ying chen, Detecting Offensive Language in Social Media to Protect Adolescent Online Safety, *ASE/IEEE International Conference on Social Computine (*2012).

[6] D. T. F. d. J. Maral Dadvar, Expert knowledge for automatic detection of bullies in social networks, *25th Benelux Conference on Artificial Intelligence (*2013).

[7] V.S. Sourabh Parime, Cyberbullying Detection and Prevention: Data Mining and Psychological Perspective, *International Conference on Circuit, Power and Computing Technologies [ICCPCT], (*2015).

[8] E.L.B.V.J.M.B.D.G.D.P.W.D.a. V. H. Cynthia Van Hee, Automatic Detection and Prevention of Cyberbullying, The First International Conference on Human and Social Analytics (2015).

[9] J.B. Sri Nandhinia, Online Social Network Bullying Detection Using Intelligence Techni-ques, International conference on Advanced Computing Technologies and Applications (2015).

[10] V.H.D.N.S. Divyashree, An Effective Approach for Cyberbullying Detection and avoidance. *International Journal of Innovative Research in Computer and Communication Engineering (*2016).

[11] B.H. Elaheh Raisi, Cyberbullying Identification Using Participant-Vocabulary Consistency, ICML Work-shop on Data4Good*:* Machine Learning in Social Good Applications (2016).

[12] Sentiment Analysis with the Naive Bayes Classifier (2016). [Online]. Available: http://atas pinar. com/ 2016/ 02/15/sentiment-analysis-with- the-naive-bayes-classifier/. [Accessed 2017].

[13] M. Vicente, F. Batista and J.P. Carvalho, Twitter gender classification using user unstructured information, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istanbul (2015).

[14] Twitter, Public Stream API, [Online]. Available: https://dev.twitter.com/streaming/public. Accessed (2017).

[15] Linguistic Inquiry and Word Count (LIWC), [Online]. Available: https://liwc.wpengine.com/.

[16] A.L. Andrei, Development and Evaluation of Tagalog Linguistic Inquiry and Word Count (LIWC) Dictionaries for Negative and Positive Emotion (2014).

[17] Mahalakshmi, K. and R. Prabhakar. Hybrid Optimization of SVM for Improved Non-Functional Requirements Classification. International Journal of Applied Engineering Research 10(.20): (2015)

[18] Mahalakshmi K, MathiVanan P, MohanaPriya D, R Keerthika, A.Nagajothi, An Optimized Support Vector Machine for Classifying Opinions in M Learning Systems Applied To Biotechnology Domain. Research Journal of Biotechnology, World Research Journals, special issue 1: 168-176 (2017)