Prediction of Cyberbullying Incident on Social Media Network

Sanmit Vartak[1], Ajinkya Vaydande[2], Jagruti Varule[3]

[1-3]*Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India*

Prof. Dnyaneshwar Dhangar[4]

[4]*Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India*

Abstract - *Cyber bullying is a intimidating and threating people on social media network via communication devices. While internet based life offer extraordinary correspondence openings, the rise of user generated content in social networking sites, Cyberbullying is also increasing concern and has gained considerable attention. It has become increasingly common among teenagers. Recent research report that cyber bullying is a growing problem among young people. Proper avoidance of profanity words depends on the identification of offensive words that requires intelligent systems to identify potential risk automatically. The aim of this work is building a predictor that can predict the occurrence of cyber bullying incidents before they happen. The main focus of this project is to automatically detect incidents of cyber bullying on social media by analyzing posts written by bullies and victims. A detail analysis of the labelled data is then presented, including a study of relationships between cyber bullying and most of features cyber aggression, profanity, social graph features, temporal commenting behavior and image content.*

Keywords— Cyberbullying Detection, Victimization, Social media, physical bullying, social bullying, Verbal bullying.

1. INTRODUCTION

As online social networks have grown in popularity, instances of cyberbullying incidents in online social network became an increasing concern in recent years. Quite half of teenagers have reported being the victims of cyberbullying. Moreover, research has found that links between experiences of cyberbullying and negative outcomes are decreased performance in school, truancy, and violent behavior, and potential devastating psychological effects like depression, low self-esteem, suicide ideation, which can have future effects within the longer-term lifetime of victims. Incidents of cyberbullying with extreme consequences like suicide are now routinely reported within the favored press. Interactive Gaming: Most gaming consoles allow people to attach and play online providing an opportunity to abuse using chats and comments. Harassment: Continuously sending cruel, offensive or threatening messages. Denigration: Exposing the secrets of an individual or gossips to wreck the reputation of an individual. Flaming: Making online arguments and using offensive language. Impersonation: Breaking into the victim's account and sending emails. Trickery: Tricking the victim into revealing sensitive information and spending it on others.

The results of cyberbullying has on its victims and it is rapidly spread among high school students, there is a need for research to know how cyberbullying occurs in social network today to those effective techniques are often developed to automatically predict cyberbullying. It reports that experts within the sector of cyberbullying could favor automatic detection of cyberbullying on social media networking sites, and propose effective follow-up techniques and methods. Our work makes an important distinction between cyber aggression and cyberbullying. Cyber aggression is defined as aggressive online behavior that uses digital media during a way that's intended to cause harm to a different person. Examples include negative content and words like profanity, slang, and abbreviations that might be utilized in negative posts like hate, fight. Particularly important within the context of cyberbullying, is that the permanent nature of the web posts, the convenience and wide distribution during which aggressive posts are often made, the problem of identifying the behavior, the power to be connected and exposed to online interaction 24/7, and therefore the growing number of potential victims. The power imbalance can combat a spread of forms including physical, social, and relational, like a user being more technologically savvy than another, a gaggle of users targeting one user, or a well- liked user targeting a less popular one. Recurrence of bullying can occur anytime or by sharing a abusive post with various people. Facebook, Twitter, YouTube. We focus on Twitter for analyzing the users reporting an experience of cyberbullying. Twitter is a media-based social network, which allows users to post and comment on images. Cyberbullying on social media can happen in several ways, including posting a humiliating image of someone else by perhaps editing the image, posting means or hateful comments, aggressive captions or hashtags, and creating fake profiles pretending to be somebody else The main goal of this project is to study cyberbullying on Twitter. To do so, we first collected a large sample of Twitter data for analyzing Cyberbullying Incidents on the Twitter Social Network. This project makes the following important contributions: In this user will first search for any query on the search bar, the search bar is integrated with twitter. The result from twitter is filtered through our model. This model will extract text and compare it to the sensitive dataset we have collected and predict. A detailed analysis of the tagged information is then conferred, together with a study of relationships between cyberbullying and cyber aggression, profanity, social graph options, temporal commenting behavior, and image content.

2. BACKGROUND

In the same spirit as natural language processing challenges tasks, e.g., misbehavior detection task of CAW 2.0 [6][9], the cyberbullying detection task is primarily focused on the content of the conversations (of the text written by the participants, both the victim and the bully), regardless the known features and characteristics of those involved. Building on some social science and psychiatry studies (see, e.g., Mishnaa et al. [7], Hinduja and Patchin [8]), [10] one hypothesizes that any cyberbullying case involves both Insult/Swear wording and Second person or Person name. We hypothesize when the association Insult/Swear wording and Person Name / Second person is validated then, the occurrence of cyberbullying case is enabled. “You are crazy” the sentence does not lead to a cyberbullying case hence it does not have any insult/swear word. The sentence “I know you are not crazy” has both second person, Insult or swear word, but it is not a cyberbullying case. In other words, the presence of the aforementioned conditions for cyberbullying case is only a necessary condition but it does not systematically entail cyberbullying because of the variety of natural language modifiers to express negation and opposition. The above few examples demonstrate the complexity of the task of identification of cyberbullying case using standard natural language processing tools, which requires investigating all the textual information of the phrase. This motivates that the ideas put forward in this paper where a combination of features will be employed to tackle the various forms of cyberbullying cases, which includes explicit evaluation of the association Swear/Insult word and Second name/Person entity.

3. RELATED WORK

Since the severity of abusive comments in social networks are known and not much work has been done to prevent users from online social media abuse. But, there is an urgent need for a better system for detecting and barring these contents online. The earlier efforts on abuse classification goes back to 2009, where Dawie Yin and his colleagues explored a context based approach[1]. They have used content features, sentiment features and context features of a comment. They used a supervised machine learning approach, Support Vector Machine (SVM) with n-grams proves to be better than previous method. Analyzing Labelled Cyberbullying Incidents on the Instagram Social Network [2] contains following features: First, an appropriate definition of cyberbullying that incorporates both frequency of negativity and imbalance of power is applied in large-scale labelling, and is differentiated from cyber aggression. Second, cyberbullying is studied in the context of a media-based social network, incorporating both images and comments in the labelling. We found that labels are mostly in agreement about what constitutes cyberbullying and cyber aggression in Instagram media sessions. Third, a detailed analysis of the distribution results of labelling of cyberbullying incidents is presented, including a correlation analysis of cyberbullying with other factors derived from images, text comments, and social network meta data. Scalability and Timely detection of cyber bullying[3] Improves scalability and timeliness for vine By 2 components dynamic, multilevel priority scheduler for improved responsiveness, and an incremental feature extraction and classification stage for scaling. This research is to propose a cyberbullying detection system with two key characteristics, namely, scalable to handle large OSNs without sacrificing accuracy and timeliness of raising an alert when a cyberbullying instance takes place. In a study by Dinakar et al [4], states that individual topic-sensitive classifiers are more effective to detect cyberbullying. They experimented on a large corpus of comments collected from Youtube.com website. Ellen Spertus [5] tried to detect the insult present in comments, they used a static dictionary approach and defined some patterns on sociolinguistic observation to build feature vector which had a disadvantage of high false-positive rate and low coverage rate.

4. PROPOSED SYSTEM PROBLEM STATEMENT

Cyberbullying is bullying that takes place in cyberspace through various mediums including online chats, text messages and e-mails. It is a big problem on social media websites like Facebook and Twitter. Cyberbullying is difficult to detect and stop due to it happening online. The problem we face is to come up with a technological approach that can help in automatic detection of bullying on social media. The approach we will explore is a system capable of automatically detecting and reporting instances of bullying on social media platforms.

5. DETAILED DESIGN

Offensive comments/posts from client will be send to our server, from our server the comment/post will send to the twitter’s server with tokens. After getting response from the twitter’s server the offensive posts will be filtered by applying various techniques on comments/post our server and the abusive free content will be displayed as a result to the twitter.

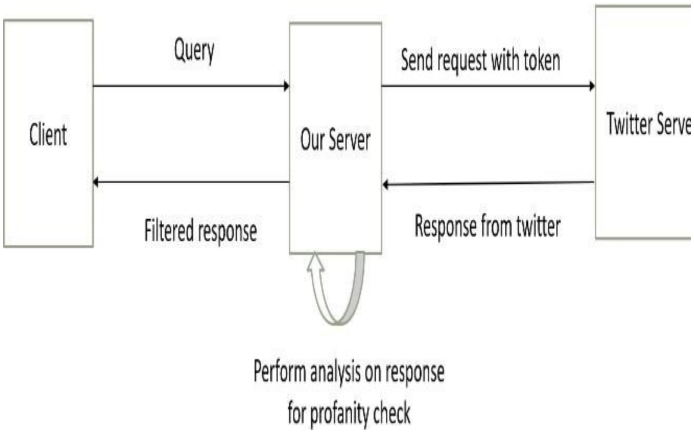


Fig A: System Design

6. WORKFLOW

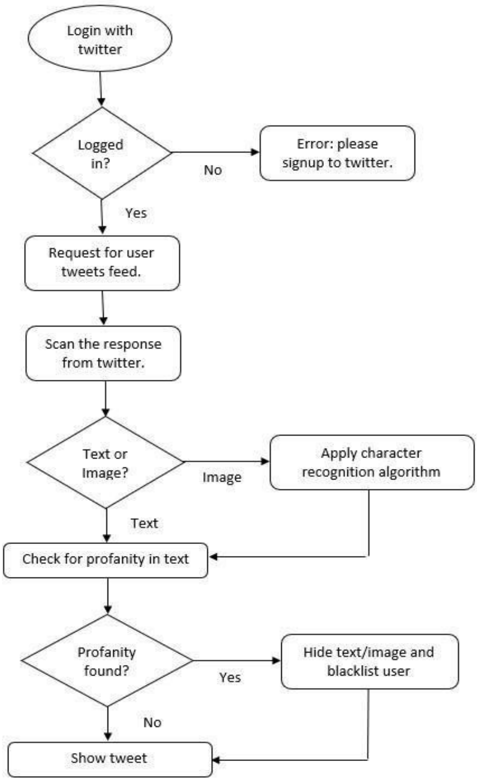

Login with twitter
Logged in?
No
Error: please signup to twitter.
Yes
Request for user tweets feed.
Scan the response from twitter.
Text or Image?
Image
Apply character recognition algorithm
Text
Check for profanity in text
Profanity found?
Yes
Hide text/image and blacklist user
No
Show tweet

Fig B: Flowchart

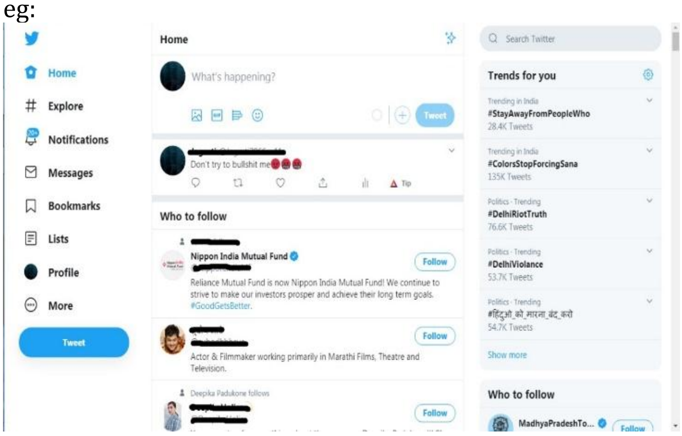
The above figure shows flow of the project. User will first login to twitter, If the user has no account then he must signup first. If login is successful user will request for tweets feed. Result will be displayed after requesting. Moderator will check for profanity words within posted images and texts by applying character recognition algorithm and display the result. If the result contains profanity in within text it will be highlighted by asterisk (*) and if found in image it will hide the image through which we will understand the respective image contains some abusive content.

7. RESULT

In our system by applying various technologies through which we got the desired result.

First of all a person (bully) tweets offensive on the social media (here we are focusing on twitter) where every twitter handler can see the offensive tweet(s)/posts. People cannot avoid this they can only do report on that particular account so that the social media team will review that particular account holder (bully's account). People can get these type of offensive posts suddenly which may make them harass, anger, depress, etc.

eg:



[redacted]
Don't try to bullshit me😡😡😡

Who to follow

[redacted]
Nippon India Mutual Fund
[redacted]
Reliance Mutual Fund is now Nippon India Mutual Fund! We continue to strive to make our investors prosper and achieve their long term goals. #GoodGetsBetter.

[redacted]
Actor & Filmmaker working primarily in Marathi Films, Theatre and Television.

Deepika Padukone follows
[redacted]

Trending in India
#StayAwayFromPeopleWho
28.4K Tweets

Trending in India
#ColorsStopForcingSana
135K Tweets

Politics · Trending
#DelhiRiotTruth
76.6K Tweets

Politics · Trending
#DelhiViolance
53.7K Tweets

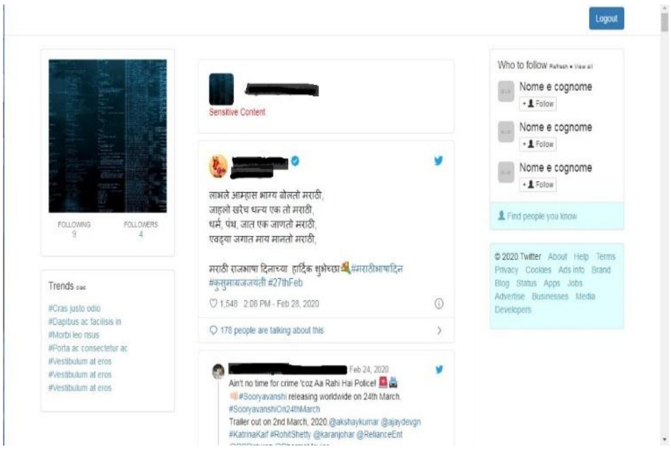
Politics · Trending
#हिंदुओं_को_मारना_बंद_करो
54.7K Tweets

MadhyaPradeshTo...

Figure1: Tweets/Post on social media

On the other hand in our system if any bully posts some tweets which is offensive it will automatically detect the posts having abusive, offensive words and will show that the post is having sensitive content. So if a sudden offensive posts appears it will not be seen by end user. Hence the offensive post will not be seen and will prevent end user from the offensive posts.

eg:


Logout
FOLLOWING
9
FOLLOWERS
4
Trends
Who to follow
Follow
Follow
Follow
Find people you know
© 2020 Twitter About Help Terms Privacy Cookies Ads info Brand Blog Status Apps Jobs Advertise Businesses Media Developers

[redacted]
Sensitive Content

[redacted]
लाभले आम्हास भाग्य बोलतो मराठी,
जाहलो खरेच धन्य एक तो मराठी,
धर्म, पंथ, जात एक जाणतो मराठी,
एवढ्या जगात माय मानतो मराठी,

मराठी राजभाषा दिनाच्या हार्दिक शुभेच्छा #मराठीभाषादिन #कुसुमाग्रजजयंती #27thFeb
1,548 2:08 PM - Feb 28, 2020
178 people are talking about this

[redacted] Feb 24, 2020
Ain't no time for crime 'coz Aa Rahi Hai Police!
#Sooryavanshi releasing worldwide on 24th March.
#SooryavanshiOn24thMarch
Trailer out on 2nd March, 2020 @akshaykumar @ajaydevgn #KatrinaKaif #RohitShetty @karanjohar @RelianceEnt

#Cras justo odio
#Dapibus ac facilisis in
#Morbi leo risus
#Porta ac consectetur ac
#Vestibulum at eros
#Vestibulum at eros
#Vestibulum at eros

Nome e cognome
Nome e cognome
Nome e cognome

Figure 2: Tweets/Post on our system

8. CONCLUSION AND FUTURE WORK

This paper tries to address the issue of cyber-bullying in media-based social network. It is an appropriate definition of cyberbullying that incorporates both frequency of negativity and imbalance of power is applied in large- scale labelling, and is differentiated from cyber aggression. This proposed model will help cyber-investigators and researchers pursuing the task of cyber-bullying detection.

Cyberbullying is studied in the context of text of a media-based social network, incorporating both images and comments in the post. It was observed that identifying the right set of keywords is an essential step for getting better results during sentiment analysis, especially during topic modelling task. A detailed analysis of the distribution results of cyberbullying incidents is presented, including analysis of cyberbullying with other factors derived from images, text comments. We found a significant number of media sessions containing profanity and cyber- aggression were consisting as cyberbullying. We observed that media sessions with very high percentage of negativity above 60-70frequent commenting. Finally, cyber- bullying has a higher probability of occurring when media sessions contain certain linguistic categories such as death, appearance, religion and sexuality content. While this paper has address the understanding of cyberbullying in a media-based mobile social network, there remain a number of areas for improvement. One theme for future work is to improve the performance of our classifier. New algorithms should be considered, such as deep learning and neural networks. More input features should be evaluated, such as new image features, mobile sensor data, etc.

9. REFRENCES

1. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards : Detection of Harassment on Web 2.0. in CAW 2.0 '09, Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain (2009)
2. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. Springer International Publishing 2015 T.-Y. Liu et al. (Eds.): SocInfo, 2015.
3. Scalable and Timely Detection of Cyberbullying in Online Social Networks. In SAC 2018: SAC 2018: Symposium on Applied Computing , April 9–13, 2018,
4. K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media,Barcelona, Spain, 2011.
5. Spertus, E., Smokey: Automatic recognition of hostile messages. In: Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence, pp. 1058–1065 (1997).
6. F. Mishna, M. Saini, and S. Solomon, Ongoing and online: Children and youth's perceptions of cyber bullying. Children Youth Services Rev. 31, 12, 1222–1228, 2009
7. S. Hinduja and J. W. Patchin, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," Youth Violence And Juvenile Justice, vol. 4, 2006, pp. 148–169.
8. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, S. Mishra. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. SocInfo 2015,pp.49-66,2015.
9. P. Burnap, M. L. Williams. Cyber Hate Speech on Twitter: An Application.
10. A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in 5th Annual ACM Web Science Conference, 2013, pp. 195–204.