A Survey on Cyberbullying

Prasanna Kumar G.
Department of Information
Science & Engg.
NIEIT, Mysuru

Sudeep J.
Department of Information
Science & Engg.
NIEIT, Mysuru

Chandru A. S
Department of Information
Science & Engg.
NIEIT, Mysuru

Abstract: **With the advancement of communication technologies and rapid growth of internet, data is generated abundantly at a complex rate. Significantly, Social networks has become one of the powerful tool for data generation and data exchange. SNS users meet other people through online community in real and virtual world in cyberspace. The rise in popularity of social networking has significantly contributed to the growth in offensive behaviors giving birth to one of the most critical problem called as cyber bullying. This paper presents survey about various approaches used for identification, detection of cyber bullying in social networks and its effects on web users.**

Keywords—SNS(social networking services). TEDAS(Twitter-based Event Detection and Analysis System), LSF (Lexical Syntactic Feature), CDE(Crime and Disaster related Events),

I. INTRODUCTION

Internet has become one of the most important useful source of information in recent years. Internet users from all over the world utilize and access varieties of social media and social network services (SNS) as a fundamental of their personal networking, relationship collaboration, transferring and sharing of knowledge within the communities. To discuss on this, further, the term “Online Social Networking” is defined as social software that has been used to develop social networks [1]. Also, the sites that

provide “Online Social Networking” services assists users in forming an impression or perception, in maintaining and acquiring new relationships in the SNS [2]. We can deduced that SNS users meet other people through online community in real and virtual world in cyberspace, allowing users to demonstrate their social networks clearly and maintain connection and networking with others.

Social media has become broadcast medium for many bloggers to broadcast the information in the form of blogging. Twitter is a micro-blogging service that evolved as a disruptive platform that is meant for the users to broadcast their daily activities, feelings and opinion by posting simple tweets (messages) within their friends network. The topics range from daily life to current activities, experience sharing, personal opinions and other interests. The social networks such as Face book, LinkedIn, Twitter, MySpace etc., has significantly embarked the way of sharing the information across the globe. Around 6.5 trillion active Twitter users [3] and is become part of their life, where everyone can share the information and opinion on, as anyone of this amount of users can't live without Twitter. Hence, we can say that micro-blogging tools such as Twitter, facilitates the sharing of one’s user short messages either publicly or within a social network, depending on the user’s privacy setting.

With the recent popularity of Twitter, it is important to know why and how people use this tool, as Twitter can sometimes used to abuse for unethical by irresponsible users to cyber bully and post something bad and harm individual’s personally. The emergence of these SNSs has caused an increase in cyber bullying circumstances, particularly among the teenagers[4]. Hence, it is important to identify the cyber bullying event and the attacking messages in social media.

Though cyber bullying might not cause any physical damage initially, however, it likely caused destructive psychological effects, like low self-esteem, mental depression, suicide consideration and even suicide [5]. A fatal cyber bullying incident had happened on MySpace SNS[6], whereby Megan Meier, a 13-year-old teen became increasingly distressed by the online harassment being directed at her, and eventually decided to end her life by hanging herself in her bedroom in 2006. Hence, recognizing the cyber bullying event itself is not efficient in combating cyber bullying per se, as we need to identify the real user of the cyber bully in order to arrest them for justice, and to prevent further similar cases to happen.

It is reported in The Star Online (2014) that a total of 389 cyber bullying reports were lodged by Internet users to the Cyber999 Help Centre in 2013 in Malaysia, which draw a 55.6% upsurge from 250 cases in 2012. Hence, by referring to this statistic, we can deduced that cyber bullying not only happened to the foreign teenagers (as mentioned earlier), it has haunted the SNS users especially in Malaysia and caught the attention of the government in addressing this social problem. However, currently there’s no any existing system that can detect the cyber bullying event based on the location of the cyber bullying event happened in our country and report the mentioned cases to police.

Thus, it’s a motivation to create a web-based application, i.e. the Cyber bullying Detection System on Twitter, with the key function to effectively discover the cyber bullying related tweets from Twitter and providing reasonable solution thereafter. With this system, the users can identify the cyber bullying related tweets based on the keywords and populate it in a news feed form. By doing this, it allows users to determine the identities of the cyber bullies and the victims from the cyber bullying tweets.

Besides that, the cyber bullying detection system is effectively useful in detecting the locations of the cyber bullies and/or the victims thru a demographic representation, by processing the captured tweets. Also, it will allow the users to generate reports to higher authorities, i.e. police reports, based on case's severity and needs.

In conclusion, with the advent of this cyber bullying detection and solution system in Twitter, it will help the authorities to monitor, regulate or at least decrease the harassing incidents in cyberspace in Malaysia. With the implementation of this system, this will also help to raise the cyber bullying awareness among the Twitter users, and posting the tweets responsibly in the social media, as posting irritating tweets is illegal and bullies can be convicted under the Computer Crimes Act, the Penal Code or the Juvenile Act, depending on the nature or severity of the case [7].

II. RELATED WORKS

The rise of social media platforms in recent years brought up huge information resources that involve new approaches to study the respective data. The social media has now gained enormous attention of the research community, as there are trying to gather, analyze and comprehend, the structure and the interconnection of the user's profile, while taking consideration of the interactions among the users' populations. This is because people nowadays utilize Social media such as Twitter not only during leisure time, but also at workplace to keep up with what's new and what's happening with one another, and people tend to spend most of their time expressing their feelings and their daily life experience and opinions through Twitter[8].

Twitter is currently one of the most popular micro blogging platforms [9]. Users interact with this system through Web interface, mobile application, instant messaging (IM) agent or sending SMS updates. The users can actually choose to make their updates or profiles public or only available to their followers (friends). There are several researches being done to investigate the usage and the communities in Twitter. Java, A.,[10], investigate the motivation of research user's in adopting this specific micro blogging platform, i.e. Twitter. As mentioned in this research, there's still a shallow studies that have been done on this form of communication and information sharing, and hence, further study on the topological and geographical structure of Twitter's social network have been carried out in this research in attempting to comprehend the user intentions and community structure in micro blogging.

Cyber bullying can be defined as a type of harassment (or bullying) that takes place online, via e-mail, text messaging, or online forums, such as social networking sites. Social networks provide ideal background for data gathering and information that might enable the criminals to execute their crime, for example, by determining one's that is a vulnerable or 'suitable' victim. We categorized these kind of crime as cyber-related crime and we are expanding its definition to include cyber bullying as one of the serious offense in cyber realm as it has resulted in death [11].

Statistical report investigated by Cyber Security Malaysia in 2007 showed that 60 cases have been reported involving cyber bullying. Although the report illustrated some isolated cases, however, the fact that this issue has already happened in many countries around the world. Not only that, based on the study by Norton Online 2010, Malaysian children spent an average of 19 hours a week on the internet [12], while the same survey also found that nine out of ten children in Malaysia has been exposed to negative experiences or element from the online use. According to the report by Cyber Security Malaysia, most cyber bullies and their victims have close contact including their close friends, ex-spouses and former colleagues. Thus, the existing problem required serious attention and solution. Cyber bullying is a serious sign and should be addressed by all parties and their concerns on the matter are necessary including parents, teachers, and the surrounding community at large.

Some previous research has discussed cyber bullying in social media. A research have been conducted to detect offensive language in social media of which incorporating a user's writing style, structure and specific cyber bullying content as features to predict the user's potentiality to send out offensive messages[13]. The technique that has been used to identify offensive language is the Lexical Syntactic Feature (LSF) approach and it is successful detecting some offensive content in social media, which has achieved precision of 98.24%, and recall of 94.34% and also succeeds in detecting users who sent offensive messages, achieving precession of 77.9%, and recall of 77.8% (Chen et al. 2012).

Besides that, another research paper proposed an architecture of a platform that automates the analysis of online social network behavior with the ultimate goal of tracing harmful content (Vanhove T, Leroux P, Wauters t, Turck F.D., 2013). This pluggable architecture made up of several components based on predetermined requirements, i.e. performance, scalability, reusability and extendibility. Analysis modules detect inappropriate content and high risk behavior after which domain services accumulate these results and flag user profiles if necessary. This platform uses text, image, audio and video based analysis modules to detect inappropriate content or any high risk behavior. With this system, the moderators of social networks will be able to quickly and accurately scan the network feed and made intervention if required[14].

With the rapid and wide coverage of Twitter, events can be discovered in an instant manner by monitoring and observing the incoming tweets. The event detection system, Twitter-based Event Detection and Analysis System (TEDAS), (R. Li, K. H. Lei,R. Khadiwala, Chang, 2012) employs an adapted information retrieval architecture that covers an online processing and an offline processing part. The offline processing is based on a fetcher accessing Twitter's API and a classifier to mark tweets as event-related or not event related. Not only have that, this system can help in identifying and examining events by exploring rich information from Twitter. From this research, there are three main functions proposed, which are detecting new events, ranking events based on their priority, and generating spatial and temporal patterns for the events detected. The TEDAS system is mainly focus on the Crime and Disaster related Events (CDE), for instance car accidents. For classifying tweets as CDE events, three features are taken into consideration, that is content features (e.g., inclusion of lexicon words), user features (e.g., number of followers), and usage features (e.g., number of retweets). Furthermore, at system level, it not only explored valuable and novel features from the Twitter, it also assist in classify and rank tweets, and predicting the locations from tweets also be made possible, as well as retrieving most of CDE tweets based on millions of tweets and users, with a set of well-defined words. The architecture of TEDAS is shown as the Figure 1 below. From this literature, we can see that it only covered the CDE related events, for which it is lacking the cyber bullying related events detection. Hence, in this research, we are going to focus on the cyber bullying detection, particularly in Twitter social media.

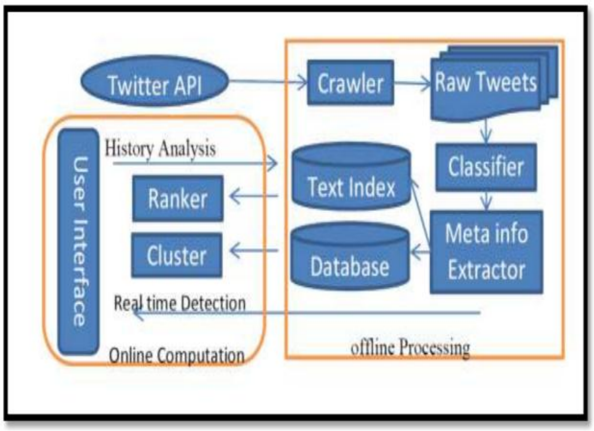

Figure 1: System Architecture of TEDAS.

Another similar event detection system, a Semi-supervised Targeted Event Detection (STED) system (Hua, T., Chen, F., Zhao, L., Lu, CT., Ramakrishnan, N., 2013) that helps users to automatically detect and interactively visualize events of targeted type from twitter, for instance, crimes, civil unrests, and disease outbreaks[15]. The STED model first applies transfer learning and label propagation to automatically generate labeled data, thereafter acquired a customized text classifier based on mini-clustering, and eventually applies fast spatial scan statistics to estimate the locations of events. With STED, a user can query for events pertaining to their specific interests and analyze its spatial and temporal features. Thereafter, target-interest variables that covers time, location, topic and keywords can be set in the system interface. Users are allowed to choose date and topic, as well the keywords in the right part of the interface. Also, the users can find the detailed information of corresponding event by clicking on one of the ballons, where it represent the tweets ranked by their relativity to users' interests. With the system proposed, STED can also possibly investigate the targeted interested events spatially and temporally, by using the historical statistics analysis interface, given a city and historical period range.

Walking through these research papers, it is promising to implement my proposed research with similar functionalities that made possible through the TEDAS and STED system. From the mentioned researches, it is possible to create a web-based system that recognize the cyber bullying tweets, identify the cyber bullying users (cyber bullies and victims), detect the locations of the victims and cyber bullies thru a demographic representation in a map feed, as well as to populate the cyber bullying tweets in my system interface.

In a recent study on cyber bullying detection, gender specific features were used and users are categorized into male and female groups. It is limited only to gender feature. In other study9, NUM and NORM features were devised by assigning a severity level to the bad words list (nosewaring.com). NUM is a count and NORM is a normalization of the bad words respectively. The dataset consisted of 3,915 posted messages crawled from the Web Site, Formspring.me. It showed only 58.5% accuracy, which is very less accuracy.

Proposed a system allowing OSN users to have a direct control on the messages posted on their walls4. This is done by using flexible rule-based system, this system allows users to customize the filtering criteria to be applied to their walls, and a Machine Learning based classifier will automatically label messages using content-based filtering. This approach is incapable of capturing more complex relationships at a deeper semantic level.

In a research work by Massachusetts Institute of Technology a system to detect cyber bullying through textual context in YouTube video comments is developed. The system classifies the comment in a range of sensitive topics such as sexuality, culture, intelligence, and physical attributes and determining what topic it is. The system shows less precise classification outcome and increased false positives. In- using a bag-of-words approach examined a baseline text mining system and improved by including sentiment and contextual features. Even with those models, a vector machine learner produce a recall level of 61.9%.

In bullying traces is identified using a variety of natural language processing techniques. Online and offline instances of bullying are traced. To identify the bullying they use sentiment analysis system and Latent Dirichlet Analysis to identify topics. In this method, the instances of bullying is not accurately detected.

Other interesting works in this area performed harassment detection from comments and chat datasets provided by a content analysis workshop (CAW). Various features were generated including: TFIDF as local features; sentiment feature, which includes second person and all other pronouns like ‘you’, ‘yourself’, ‘him’, ‘himself’ and foul words; and contextual features. Increased false positive is its limitation. Research on online sexual predators detection associate the theory of communication and text-mining methods to analyze difference between predator and victim conversations, as applied to one-to-one communication such as in a chat-log dataset. The as usual methods are based on the keywords. It involves high semantic and contextual work.

Generally most existing systems are focusing on effects after cyber bullying incident and there is no system for online cyber bullying detection. Intelligence techniques are also not used in cyber bully detection. The proposed system is to detect the cyber bullying activities and classify them as Flaming, Harassment, Racism and Terrorism, which helps to prevent the cyber bullying victims from facing effects of cyber bullying and take necessary actions like blocking, law enforcement or taking corresponding legal actions accordingly.

III. CONCLUSION

Until now, very few implementation has been done to identify and detect the activities of cyber bullying on social networks. Cyber bullying has become a menace in social networks and it requires extensive research for identification and detection over web users. We have proposed a brief summary about various approaches used for identification, detection of cyber bullying in social networks and its effects on web users.

REFERENCE

[1] L.Q.L. Mew, “Online social networking: a task-person-technology fit perspective”, PhD dissertation, School of Business, George Washington University, 2009.

[2] S. Tom Tong, et al., “Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook”, Journal of Computer-Mediated Communication, 13(3), 2008, pp. 531–549, April 2008.

[3] Statistic Brain. (2014). Twitter Statistics. Retrieved from http://www.statisticbrain.com/twitter-statistics/

[4] Livingstone, Sonia and Bober, Magdalena (2004) UK children go online: surveying the experiences of young people and their parents. 2. London School of Economics and Political Science, London, UK. (Livingstone et al., 2004).

[5] S. Hinduja and J. W. Patchin (2010). “Cyberbullying research summary, cyberbullying and suicide,”.

[6] Tavani, Herman. T. (2013). Introduction to Cyberethics: Concepts, Perspectives, and Methodological Frameworks? In H. T. Tavani, Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing. River University – Fourth Edition: Wiley, pp.1-2.

[7] Anandarajah, Anita (2004, September 30) COVER STORY: Cyber bully. New StraitsTimes. Retrieved from: http://www.cybersecurity.my/en/knowledge_bank/news/2004/main/detail/904/index.html

[8] Zhao, D. & Rosson, M.B., (n.d). 2008. How Might Microblogs Support Collaborative Work? Retrieved from http://research.ihost.com/cscw08socialnetworkinginorgs/papers/zhao_cscw08_workshop.pdf

[9] Twitter (March 21, 2012). "Twitter turns six". Twitter. Retrieved from: https://blog.twitter.com/2012/twitter-turns-six

[10] Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, WebKDD/SNA-KDD ’07 (pp. 56–65). New York, NY, USA: ACM.

[11] Tavani, Herman. T. (2013). Introduction to Cyberethics: Concepts, Perspectives, and Methodological Frameworks? In H. T. Tavani, Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing. River University - Fourth Edition:Wiley, pp.1-2[12] Utusan Malaysia. (2010/2011). ’Mangsa buli di laman sosial’.

[12] Y. Chen, Y. Zhou, S. Zhu and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, Amsterdam, 2012, pp. 71-80.

[13] T. Vanhove, P. Leroux, T. Wauters and F. De Turck, "Towards the design of a platform for abuse detection in OSNs using multimedial data analysis," *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, Ghent, 2013, pp. 1195-1198.

[14] Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. 2013. STED: semi-supervised targeted-interest event detectionin in twitter. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '13), Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy (Eds.). ACM, New York, NY, USA, 1466-1469. . DOI: http://dx.doi.org/10.1145/2487575.2487712