

# Cyberbullying Detection System on Twitter

**Liew Choong Hon, Kasturi Dewi Varathan**

*Department of Information Systems,*

*Faculty of Computer Science and Information Technology*

*University of Malaya, Malaysia.*

*kelvinliew1991@hotmail.com, kasturi@um.edu.my*

## Abstract

*Social networks such as Twitter is a microblogging service and broadcast medium that evolved as a disruptive platform to serve the purposes for the users to broadcast their daily activities, feelings and opinion by posting simple tweets (messages) within their friends circle. Cyberbullying is a type of harassment that takes place social networking sites which enable the cyberbullies to execute their crime on the vulnerable victims and this serious offense in cyber world has resulted in death. Hence, the Cyberbullying Detection System on Twitter is a solution with the aim to effectively discover the cyberbullying related tweets from Twitter. The system will be used by the organizations member to monitor the social network community's activities, especially the cyberbullies and victims in Twitter and hence preventing the cyberbullying event to be deteriorated. In order to accomplish these objectives, 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. In our research, we have specified the cyberbullying related keywords from the research experts, and captured the targeted tweets.*

**Keywords:** *Twitter; cyberbullying detection system; cyberbullies; victims*

## 1. Introduction

To date, 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 further discuss on this, firstly, the term "Online Social Networking" is defined as social software that has been used to develop social networks (L.Q.L. Mew, 2009). 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 (S. Tom Tong, et al, 2008). We can deduced that the online community is a place for the SNS users to meet people allow users to meet other people in the other sides of the world in cyberspace, allowing users to demonstrate their social networks clearly and maintain connection and networking with others, in real world and virtual world.

Social networks such as microblogging is a broadcast medium that exists in the form of blogging. Twitter is a microblogging 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 circle. The topics range from daily life to current activities, personal

opinions, experience sharing and other interests. The social networks, i.e. Twitter has turned into something that is indivisible to these 645,750,000 active Twitter users (Statistic Brain, 2014) and becoming 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 see that microblogging tools, i.e. 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 these tool, as Twitter can sometimes used to abuse for unethical by irresponsible users to cyberbully and post something bad and harm individual's personally. The emergence of these SNSs has caused an increase in cyberbullying circumstances, particularly among the teenagers (Livingstone et al., 2004). Hence, it is important to identify the cyberbullying event and the attacking messages in social media.

Though cyberbullying 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 (S. Hinduja and J. W. Patchin, 2010). A fatal cyberbullying incident had happened on MySpace SNS (Tavani, Herman. T. , 2013), 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 cyberbullying event itself is not efficient in combating cyberbullying per se, as we need to identify the real user of the cyberbully 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 cyberbullying 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 cyberbullying 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 cyberbullying event based on the location of the cyberbullying 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 Cyberbullying Detection System on Twitter, with the key function to effectively discover the cyberbullying related tweets from Twitter and providing reasonable solution thereafter. With this system, the users can identify the cyberbullying 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 cyberbullies and the victims from the cyberbullying tweets.

Besides that, the cyberbullying detection system is effectively useful in detecting the locations of the cyberbullies 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 cyberbullying 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 cyberbullying 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 (Anandarajah, Anita, 2004).

## **2. 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 (Zhao, D. & Rosson, M.B., 2008).

Twitter is currently one of the most popular microblogging platforms (Twitter, 2012). 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., et al., 2007), investigate the motivation of research user's in adopting this specific microblogging 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 microblogging.

Cyberbullying 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 cyberbullying as one of the serious offense in cyber realm as it has resulted in death (Tavani, Herman T., 2013).

Statistical report investigated by Cyber Security Malaysia in 2007 showed that 60 cases have been reported involving cyberbullying. 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 (Utusan Malaysia, 2010/2011), 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 cyberbullies and their victims have close contact including their close friends, ex-spouses and former colleagues. Thus, the existing problem required serious attention and solution. Cyberbullying 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 cyberbullying content as features to predict the user's potentiality to send out offensive messages. 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 behaviour 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 extendability. Analysis modules detect inappropriate content and high risk behaviour 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 behaviour. With this system, the moderators of social networks will be able to quickly and accurately scan the network feed and made intervention if required.

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 cyberbullying related events detection. Hence, in this research, we are going to focus on the cyberbullying detection, particularly in Twitter social media.

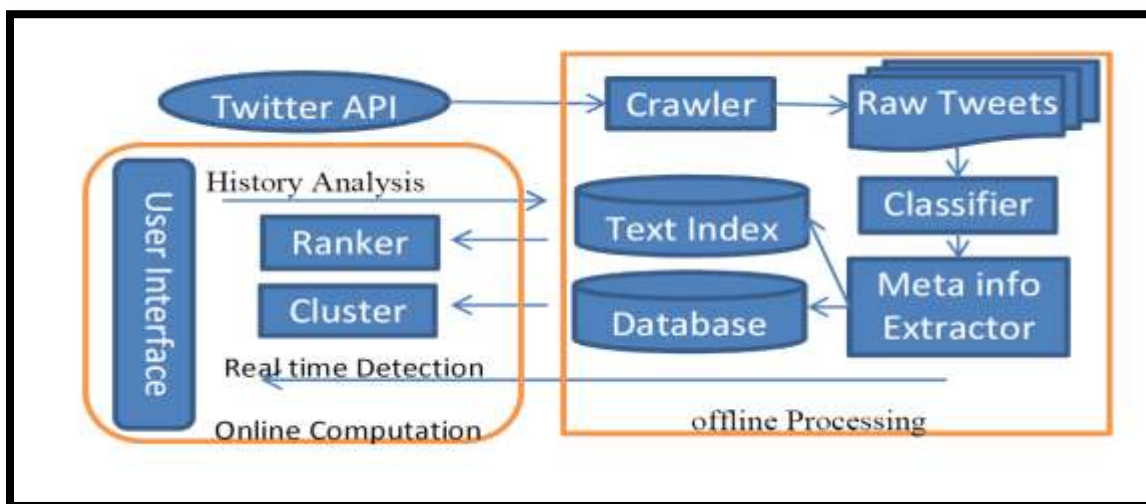


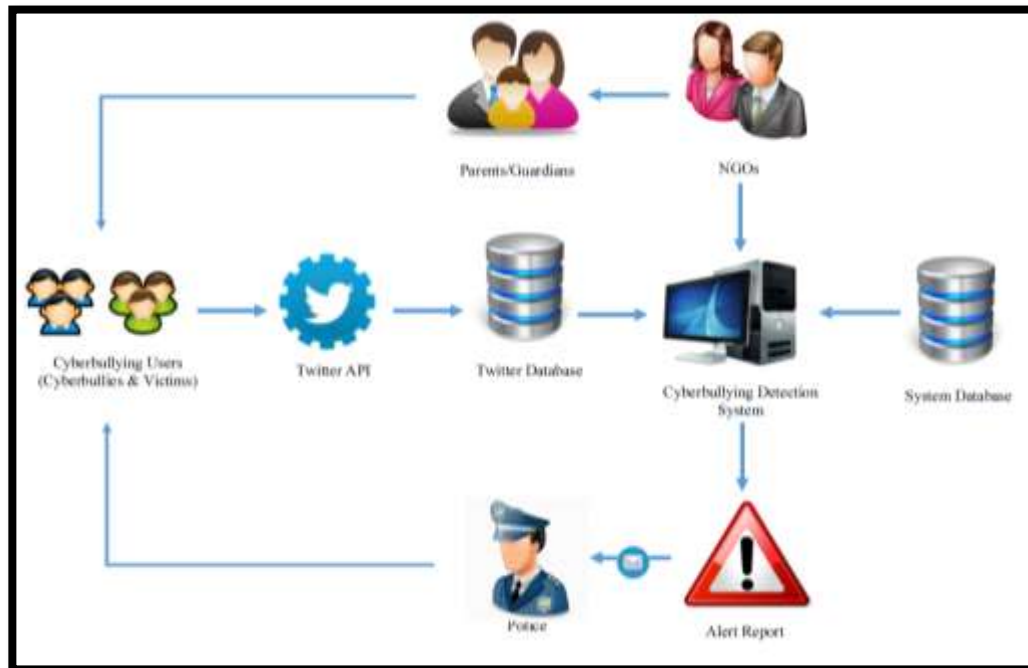
Figure 1: System Architecture of TEDAS.

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. 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 cyberbullying tweets, identify the cyberbullying users (cyberbullies and victims), detect the locations of the victims and cyberbullies thru a demographic representation in a map feed, as well as to populate the cyberbullying tweets in my system interface.

### **3. Methodology**

The implementation of the Cyberbullying Detection System on Twitter is based on PHP and HTML with the MySQL and Twitter API. This system will detect cyberbullying related tweets that have matching keywords from the database. The conceptual model of cyberbullying detection system is shown in Figure 2.



*Figure 2: The Conceptual Model of Cyberbullying Detection System on Twitter*

The conceptual model for this system describes the overall process on how the cyberbullying-related tweets can be recognized and alerts the NGOs (the user of the system), and hence alerts the nearby police station via email reports, also alert the cyberbullying user's parents or guardians in monitoring their cyberbullying activities. Initially, the user need to login into the system. The system has been coded with the OAuth token retrieved from the Twitter Developers account from the website. The system basically connect to Twitter once login is successful.

By utilizing the Twitter APIs, the cyberbullying related tweets will be retrieved by the connection of the APIs and the database that matched based on the cyberbullying keywords identified, and the cyberbullying users (cyberbully and/or victims). The system first fetches matched tweets with the cyberbullying tweets and words, also the cyberbullying users' information from the database. The results are then sent to the User Interface (UI) which will be populating the information of the cyberbullying users and the tweets itself.

Thereafter, the users (NGOs) can interact with the cyberbullying users by giving advice, warning, or counselling to monitor the cyberbullying activities. Also, the users can view and access

the tweet's location in a map form that will allow the users to identify their location more precisely. This will help in generating the topography and statistics of cyberbullying event based on specific location.

#### **4. Project Scope**

In this Cyberbullying Detection System on Twitter, we have to clarify the scope in order to accomplish this project efficaciously. Since this is a text-based cyberbullying tweets mining system, we have to clearly choose the type of text or language will be captured for this system to process. In this system, we only focus on English language with proper and formal text. Thus, the informal, short form and abbreviation text and other language text which enlightening to the cyberbullying means will not be captured in the system.

Furthermore, since the system only focusing on the text posted by the Twitter users, the punctuation and emotional icons will not be taken into account. The thing we focus in this study is only text written by users. Besides that, based on the research (S. K. H. Sanchez, 2012), we are focusing on five types of cyberbullying related word that we deduced from the study, that include: 'gay', 'bitch', 'slag', 'homo', 'dike', 'queer', 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 SNS. 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 victims, and their location.

#### **5. Results and Discussion**

From this research, a cyberbullying detection system will be developed. By accessing this system, the users, i.e. NGOs can identified the cyberbullying-related tweets that has been retrieved by the connection of the APIs and the database which contains the varieties of cyberbullying related keywords mentioned earlier on. The system will fetch the matched tweets with the cyberbullying-related words, and the basic profile information of cyberbullying users (cyberbully and/or victims) from the database. These results will be sent to the interface which will be showing and populating the information of the cyberbullying users and the tweets in a Timeline. Figure 3 shows the screenshot of the prototype developed on detecting a cyberbullying tweets via specified date and particular location.

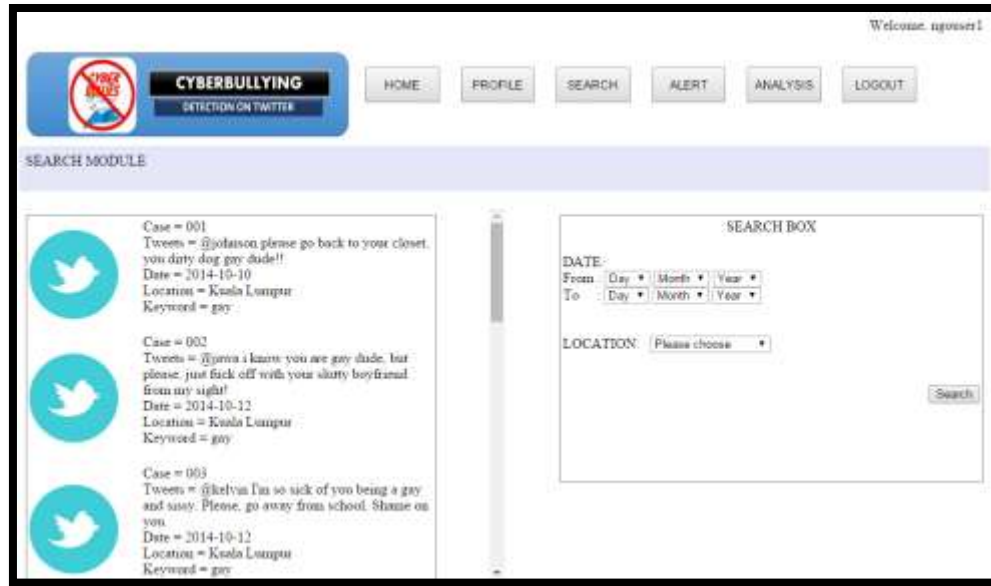


Figure 3: Cyberbullying tweets detection interface.

Next, the user will be notified and alert on the cyberbullying event, and will be able to see the cyberbullying users profile of the users, which contain their basic details, e.g. Name, Age, Phone No, Email, Address, etc. The users will be able to access and use these information to further contact the cyberbullying users and give proper guidance and counselling for the needs. Figure 4 shows the screenshot of the system developed on populating the basic details of the cyberbullying event and the cyberbullying users, i.e. cyberbullies and victims.

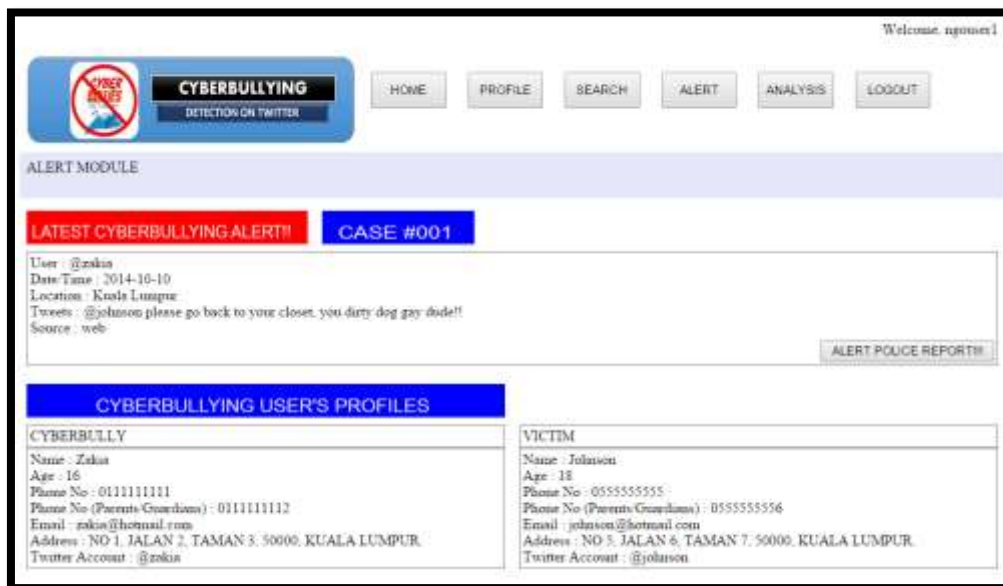


Figure 4: Cyberbullying event and users profiles.

From this information, the users can directly call the cyberbullying users by their phone number stated at the alert page, or generate a police report and email to the nearby police station



for further investigation, if the event is out of regulation. Figure 5 shows the prototype developed in generating the police report summary in reporting the cyberbullying cases.

Welcome, ngouser1

**CYBERBULLYING**  
DETECTION ON TWITTER

HOME PROFILE SEARCH ALERT ANALYSIS LOGOUT

**ALERT MODULE**

Police Report have been Generated and Sent to the destination: Police Station Bukit Bintang !!!

**POLICE REPORT SUMMARY** **CASE #001**

Generated by : ngouser1  
Date/Time : 2015-02-17 12:45:28  
Case Details : Cyberbullying Detection on Twitter  
Actions Taken : Advices and counselings have had given individually. Subjects refused to change their attitudes. Cases are required to escalate to higher authorities by lodging a report to Police Station.

**CYBERBULLYING USER'S PROFILES**

CYBERBULLY	VICTIM
<p>Name : Zakia Age : 16 Phone No : 0111111111 Phone No (Parents/Guardians) : 0111111112 Email : zakia@hotmail.com Address : NO 1, JALAN 2, TAMAN 3, 50000, KUALA LUMPUR Twitter Account : @zakia</p>	<p>Name : Johnson Age : 18 Phone No : 0555555555 Phone No (Parents/Guardians) : 0555555556 Email : johnson@hotmail.com Address : NO 5, JALAN 6, TAMAN 7, 50000, KUALA LUMPUR Twitter Account : @johnson</p>

Figure 5: Cyberbullying event police report summary.

Besides that, based on the cyberbullying users information captured, the system will further analyzed this information and populate it into a relationship table through identifying the number of victims in respond with each cyberbullies, and the locations of the cyberbullies and victims respectively. Hence, user will be able to identify the cyberbullying case statistics of the cyberbullying tweets event based on the location populated from the cyberbullying tweets. Figure 6 shows the relationship tables between the cyberbullies and the responding victims, based on their basic information consist of Name, Twitter ID, Location and the number of resulting victims of each cyberbullies.





Welcome, ngouser1

**CYBERBULLYING**  
DETECTION ON TWITTER

HOME PROFILE SEARCH ALERT ANALYSIS LOGOUT

**CYBERBULLYING ANALYSIS MODULE**

Please choose  Search

Cyberbully's Details	Cyberbully's Location	Victim's Details	Victim's Location
 <p>Name = Zakia Twitter ID = @zakia Total Number of Victims = 4</p>	Kuala Lumpur	 <p>Name = Johnson Twitter ID = @johnson Cyberbully Keyword = gay</p>	Kuala Lumpur
		 <p>Name = Jaws Twitter ID = @jaws Cyberbully Keyword = gay</p>	Kuala Lumpur
		 <p>Name = Sarah Twitter ID = @sarah Cyberbully Keyword = alke</p>	Kuala Lumpur

*Figure 6: The relationship table and statistics analyzed based on the cyberbullying users location and basic details.*

By implementing this system, the cyberbullying users will be more alert when posting a status in their profile or someone timeline. As the cyberbullying activities are monitored by the users, the cyberbullies will be aware of the consequences upon the usage of social networks, and behave in a proper manner during communication. With this system, it will instill a cyberbullying awareness within the social network communities.

## **6. Future Work and Conclusion**

The prototype has been developed based on the Cyberbullying Detection System on Twitter conceptual model resulted from the related works review and methodology development. This development of cyberbullying detection system in the social media, Twitter is an innovative idea in this research field. Our developed system not only will be able to identify the cyberbullying event, it also will allowed the users to identify the identities and information of the cyberbullying users (cyberbullies and victims), their location analysis, and generates police reports. However, the system is not fully developed in a complete stage as several scopes and limitation are not covered, including the no punctuations sensitive, and not emoticon icons sensitive and so on. To conclude, we hope that in near future, we will diverged the perspective and take those limitation mentioned into account in promoting and motivating the system results' accuracy and effectiveness. In future works, we would like to include the Malay Language cyberbullying-related keywords in our database to capture the cyberbullying tweets especially in Malaysia, to further suit the cyberbullying event detection in Malaysia context. And we hope that by the implementation of this system, the cyberbullying awareness can be raised across the country and across the social platform, i.e. Twitter, which is extensively used by the communities today.

## **7. Acknowledgements**

We would like to take this opportunity to thank University of Malaya UMRG (RP004B-13ICT) for funding this Research.

## **References:**

- [1] Anandarajah, Anita (2004, September 30) COVER STORY: Cyber bully. New Straits Times. Retrieved from: [http://www.cybersecurity.my/en/knowledge\\_bank/news/2004/main/detail/904/index.html](http://www.cybersecurity.my/en/knowledge_bank/news/2004/main/detail/904/index.html)
- [2] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012): '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 Confernece on Social Computing (SocialCom), pp. 71-80.
- [3] Hua, T., Chen, F., Zhao, L., Lu, CT., Ramakrishnan, N., "STED: Semi-Supervised Targeted Event Detection," Proceedings of the 19th ACM SIGKDD," Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data (ACM-KDD), Demo Track, 2013 (Accepted)
- [4] 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.

- [5] L.Q.L. Mew, "Online social networking: a task-person-technology fit perspective", PhD dissertation, School of Business, George Washington University, 2009.
- [6] 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)
- [7] R. Li, K. H. Lei, R. Khadiwala, Chang, "TEDAS: A Twitter-based Event Detection and Analysis System," icde, pp.1273-1276, 2012 IEEE 28th International Conference on Data Engineering, 2012.
- [8] S. Hinduja and J. W. Patchin (2010). "Cyberbullying research summary, cyberbullying and suicide,".
- [9] S. K. H. Sanchez, "Twitter bullying detection," ser. NSDI'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 15-15.
- [10] 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.
- [11] Statistic Brain. (2014). Twitter Statistics. Retrieved from <http://www.statisticbrain.com/twitter-statistics/>
- [12] 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
- [13] The Star Online. (2014). Cyber-bullying reports up 55.6% in 2013. Retrieved from: <http://www.thestar.com.my/News/Nation/2014/02/24/Cyber-bullying-u>
- [14] Twitter (March 21, 2012). "Twitter turns six". Twitter. Retrieved from: <https://blog.twitter.com/2012/twitter-turns-six>
- [15] Utusan Malaysia. (2010/2011). 'Mangsa buli di laman sosial'.
- [16] Vanhove T, Leroux P, Wauters t, Turck F.D. (2013). Towards the design of a platform for abuse detection in OSNs using multimedial data analysis. IM 2013: 1195-1198
- [17] Zhao, D. & Rosson, M.B., (n.d). 2008. How Might Microblogs Support Collaborative Work? Retrieved from [http://research.ihost.com/cscw08-socialnetworkinginorgs/papers/zhao\\_cscw08\\_workshop.pdf](http://research.ihost.com/cscw08-socialnetworkinginorgs/papers/zhao_cscw08_workshop.pdf)