Effective Approaches for Automated Textual based Cyberbullying Detection: A Survey

Josmi Jose[1] Prof. Ajith John Varghese[2]
[1]M. Tech Student [2]Guide
[1,2]Department of Computer Science & Engineering
[1,2]Musaliar College of Engineering and Technology, Kerala, 689653, India

Abstract— Online social networking sites are being rapidly increased in recent years. Through these social sites people all over the world can connect and they express their own feelings, emotions, interests etc. As the growth of online social networks, security is to be an important concern. Cyber crime is one of the major areas which committed in internet. Social Medias such as Facebook, Twitter are gaining more popularity as spreading messages to others. Messaging that provides opportunities to create harmful activities, named as cyberbullying. As the amount of crimes growing every day, it is impossible to perform manual detection on the dataset and extract useful information. Therefore, machine learning techniques are used, which has the ability to monitor and automatically detecting harmful online activities such as bullying messages, and helps to construct a healthy and safe social media environment. In this paper, we present a systematic review of published researches based on the automatic text-based content cyberbullying detection approaches. This paper essentially serves as a resource for researchers to determine where to best direct their future research efforts in this field.

Key words: Cyber Crime, Machine Learning, Social Medias, Text-based Cyberbullying Detection

I. INTRODUCTION

The social networking sites are growing rapidly for sharing information, thus users to get in close contact with others without considering cyber security. These kinds of communication may lead to some hazardous outcomes in terms of security attacks in social networks. One of the security attacks is by posting messages which contain the sharing of some kinds of abusive contents, which can emerge the threats like cyberbullying [1], [2]. Due to the recent growth of social media platforms such as Facebook and Twitter, cyberbullying is becoming more and more prevalent issue.

Cyberbullying is a new form of bullying method and is an increasingly important and serious social problem, which can negatively affect the individuals. It is defined as an aggressive, intentional actions performed by an individual or group of people via digital communication methods [3]-[5]. It is also to be known as 'Electronic Bullying' & 'Online Social Cruelty'. Fig1. Shows that the various categories of cyberbullying that generally occurs on the social networks. The victims of cyberbullying often suffer from various mental issues, ranging from depression, loneliness, anxiety to low self –esteem. The implementation for monitoring social networking sites to detect cyberbullying activities is less and far more challenging task.

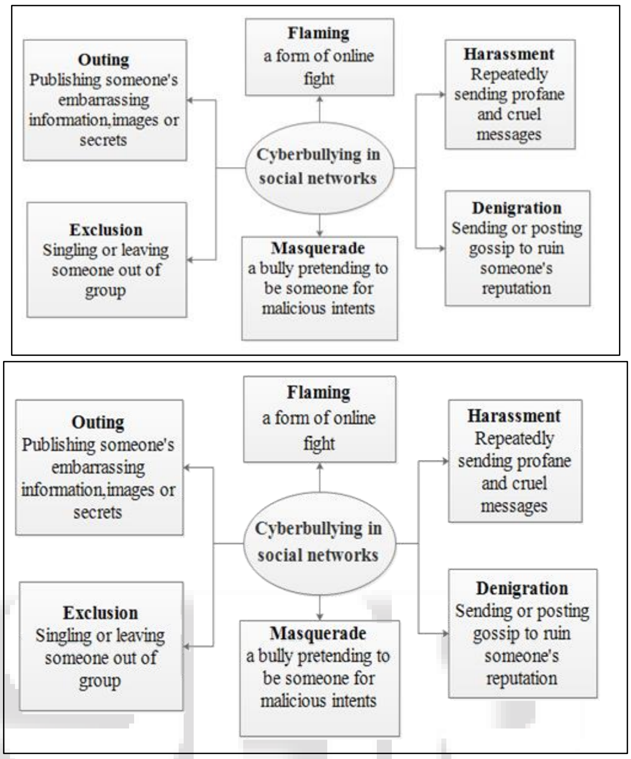
Flaming
a form of online fight
Outing
Publishing someone's embarrassing information,images or secrets
Harassment
Repeatedly sending profane and cruel messages
Cyberbullying in social networks
Exclusion
Singling or leaving someone out of group
Masquerade
a bully pretending to be someone for malicious intents
Denigration
Sending or posting gossip to ruin someone's reputation

Fig. 1: Various categories of cyberbullying in social networks

In earlier days, manual detection is considered as the most accurate detection. But it is hardly employed, because it is time consuming and labor consuming, so that the detection of cyberbullying is therefore still challenging. Data mining has been studying for decades trying to get useful information out of from a huge amount of data. So implements an automatic system for to detecting the cyberbullies. The automatic detection of cyberbullying posts is becoming an increasingly important area of research among social media researchers. And the ultimate goal of automatic cyberbullying detection is to reduce manual monitoring efforts on social media. In this paper, we spotlights on the text-based content cyberbullying detection, one of the main forms of cyberbullying.

The main objective of this survey is to gain insight into the various techniques that adopts to detect the text-based cyberbullies in online social networking sites.The rest of the paper is organized as follows .Section II presents related researches on techniques used in automatic cyberbullying detection system. Section III presents the conclusion.

II. LITERATURE REVIEW

There are numerous evidences showing that messaging in social networking sites can introduce a problem called cyberbullying[1], [2]. As insults, rumors and misinformation can be immediately disseminated to a large audience, cyberbullying in social networks is painful for victims. So successful detection of cyberbullying is one of the key importance to identify possibly threatening messages. For humans it has become very difficult to keep track of all conversations which produced online. In order to manage this amount of information in an efficient way, there is a need for intelligent techniques to signal the harmful and abusive contents automatically. This section reviews that the relevant previous works or researches related to the text representations and the automatic detection of cyberbullying.

A. Survey on Text Representation Learning

In the text-based cyberbullying detection, one of the main critical steps is the numerical representation learning for text messages. The representation learning of text is extensively studied in text mining. In text mining, natural language processing (NLP) and the information retrieval, effective numerical representation of linguistic units is a key issue. BoW model is one of the text representation methods. This model represents document in a textual collection, uses vector of real numbers which indicates occurrences of words in that document [1]. Another text representation model is topic model. Topic models which including Latent Dirichlet Allocation (LDA) , Latent Semantic Analysis (LSA) [10], [11]. The topic models are trying to define the generation of each word occurs in the document.

B. Survey on Cyberbullying Detection

In recent years, the machine learning methods are gaining increased popularity and the computational study of cyberbullying which attracted the interest of researchers. Several research area such as topic detection and affective analysis are closely related to cyberbullying detection. Because of their efforts, automatic cyberbullying detection is becoming possible.

Xu et.al presented several off-the-shelf Natural Language Processing solutions including Bag of Words (BoW) models, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) for the representation learning to capturing the bullying signals in social media [1], [6]. But they did not develop specialized models for the cyberbullying detection.

Dadvar et. al used gender- specific features for the detection of cyberbullying from MySpace dataset . They trained a SVM classifier on separate female and male authored posts. It results an improved performance over N-grams, TF-IDF, and foul words frequency, as feature sets [3]. In the follow up work they applied a hybrid approach combining supervised machine learning models with an expert system to recognize cyberbullying [7]. Dadvar et. al have shown improved performance combining user information and expert views.

Yin et.al used content information for the detection of harassing posts. They applied a supervised machine learning approach for detecting harassment. They determined the text mining system by using a bag of words approach , which was used sentiment, content and the contextual features for to train a SVM classifier for determine whether it is a harassing post or not [3], [8]. The results shows that 78% of accuracy, by recording the percentage of insult or abusive words in a post.

Dinakar et.al used the label specific features, learned by Linear Discriminative Analysis. The performance of label specific features depends on the training corpus size. In addition they construct bullyspace knowledge to improve the performance of NLP methods. Also, they deconstructed the detection of cyberbullies into the sensitive-topic detection, including sexuality, race, intelligence etc [9]. They demonstrated improved performance of label- specific classifier.

III. CONCLUSION

A number of life threatening cyberbullying experiences on social networking sites among people, especially to teenagers and adolescents have been reported internationally, this emerges a negative impact. So cyber security is becoming an important concern with increase in the use of social networking sites and internet. To provide more cyber security, implements automatic detection of cyberbullies. In this paper, we presented a survey on the various techniques and methods that available for the detection and prevention of cyber harassment.

ACKNOWLEDGEMENT

I wish to express my deepest sense of gratitude for immeasurably valuable guidance and support I have received from my guide Mr. Ajith John Varghese, MCET, PTA who took great care for providing support and facilities needed to accomplish the work. I express my deep gratitude to Almighty God for giving me the strength. Last, but not least I express my sincere thanks to my parents for their love, inspiration and moral support.

REFERENCES

[1] Rui Zhao, Kezhi Mao, “Cyberbullying Detection based on Semantic- Enhanced Marginalized Denoising Auto-Encoder”, IEEE Transactions, in press.

[2] Walisa Romsaiyud , Piyaporn Nurarak, Pirom Konglerd , “ Automated Cyberbullying Detection using Clustering Appearance Patterns” , IEEE Transaction, 2017.

[3] Vinita Nahar, Xue Li, Chaoyi Pang, Yang Zhang , “Cyberbullying Detection based on Text-Stream Classification”, in Proceedings of 11-th Australian Data Mining Conference, vol. 146, 2013.

[4] P.K. Smith, J. Mahadavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, “Cyberbullying: Its Nature and Impact in Secondary School Pupils”, Journal of Child Psychology & Psychiatry, vol. 49 , pp.376-385,2008.

[5] Krishna B. Kansara and Narendra M. Shekokar, “A Framework for Cyberbullying Detection in Social Network”, International Journal of Current Engineering and Technology, vol. 5, 2015.

[6] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, “ Learning from Bullying Traces in Social Media”, in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, pp. 656-666, 2012.

[7] M. Dadvar, D. Trieschnigg , and F. de Jong, Experts and Machines against Bullies: A Hybrid Approach to detect

cyberbullies. Springer International Publishing, pp. 275-281, Jan 2014.

[8] Yin, D. Xue, Z. Hong, L. Davisoni, B.D. Kontostathis, A. & Edwards, “Detection of Harrassment on Web 2.0, in ‘Content Analysis in the Web 2.0 Workshop’,2009.

[9] K. Dinakar, R. Reichart, and H. Lieberman, “Modelling the Detection of Textual Cyberbullying”, in the Social Mobile Web, 2011.

[10] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis”, Machine Learning, vol. 42, pp. 177-196, 2001.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”, The Journal of Machine Learning research ,vol. 3, pp. 993-1022, 2003.