Cyberbullying Detection: An Overview

Conference Paper · November 2018

DOI: 10.1109/CR.2018.8626869

3 authors:

Wan Noor Hamiza Wan Ali
Universiti Kebangsaan Malaysia
1 PUBLICATION **2** CITATIONS
SEE PROFILE

Masnizah Mohd
Universiti Kebangsaan Malaysia
45 PUBLICATIONS **90** CITATIONS
SEE PROFILE

Fariza Fauzi
Universiti Kebangsaan Malaysia
22 PUBLICATIONS **146** CITATIONS
SEE PROFILE

Cyberbullying Detection: An Overview

Wan Noor Hamiza Wan Ali
Centre for Cyber Security
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor
p93244@siswa.ukm.edu.my

Masnizah Mohd
Centre for Cyber Security
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor
masnizah.mohd@ukm.edu.my

Fariza Fauzi
Centre for Cyber Security
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor
fariza.fauzi@ukm.edu.my

***Abstract*— This paper is an overview of cyberbullying which occurs mostly on social networking sites and issues and challenges in detecting cyberbullying. The topic presented in this paper starts with an introduction on cyberbullying: definition, categories and roles. Then, in the discussion of cyberbullying detection, available data sources, features and classification techniques used are reviewed. Natural Language Processing (NLP) and machine learning are the famous approaches used to identify bullying keywords within the corpus. Finally, issues and challenges in cyberbullying detection are highlighted and discussed.**

Keywords— cyberbullying, social media, *Twitter*, detection

I. INTRODUCTION

To date, people all over the world utilize internet as a tool for communication amongst them. Online tools such as social networking sites (SNSs) are the most popular socializing tool especially for adolescents as SNSs tightly integrated in their daily practices since it can be a medium for users to interact with each other without any limitation of time or distance [1]. Nevertheless, SNSs can give negative consequences if users misuse them and one of the common negative activities that occurs in SNSs is cyberbullying which is the focus of this paper.

Cyberbullying involves a person doing threatening act, harassment, etc. towards another person. Meaning of cyberbullying is a group(s) or an individual(s) of peoples that adopt telecommunication advantages to intimidate other persons on the communication networks [2]. However, most of the researchers in cyberbullying field take into account definition of cyberbullying from [3]. According to [3], definition of cyberbullying formulated as "willful and repeated harm inflicted through the medium of electronic text".

Cyberbullying, can takes into a few forms: flaming, harassment, denigration, impersonation, outing, boycott and cyberstalking [4]. The most severe type of cyberbullying is flaming and the less severe is cyberstalking as stated in [5]. Flaming occurs between two or more individuals that argue on some incidents that involve rude, offensive and vulgar language and occurred within electronic message [6]. Flaming is the most severe type of cyberbullying because if online fight between internet's users take part, it could be difficult to recognize cyberbully and victim on that time [7]. Harassment occurs repeatedly sending of harmful message to a victim [6]. Denigration is posting about victim that untrue, rumors or cruel [6]. Impersonation happens when cyberbully disguises into a target and post bad information about that particular target with intention to bullying the target [6]. Outing occurs when cyberbully share victim's secrets or private information which can embarrassing victim [4]. Boycott is exclude a person within social interaction in social media with a purpose [8], [9]. Willard mentioned cyberstalking occurs when cyberbully send harmful messages repeatedly [6]. The cyberstalking is less severity than other categories since cyberbully (cyberstalker) could be detected directly once they send annoying messages towards victim.

The main roles involved in cyberbullying occurrences are cyberbully and victim. Given the aforementioned types of cyberbullying, there are various reasons why it happens. Apart from cyberbully and victim presences, proliferation of other roles may accentuate. According [10], they were classified the role of bullying into eight roles. These are of bully, victim, bystander, assistant, defender, reporter, accuser and reinforcer.

II. CYBERBULLYING DETECTION

A. Data Source

Cyberbullying takes place in several platforms likes text messages, instant messages, social media and online games. As reported in statisticbrain.com, the most common platforms where cyberbullying occurs is within social media with the highest ranked was *Facebook* [11].

Based on [12], authors evaluated data from *YouTube* and *Formspring*. 50% from *YouTube* dataset were allocated as training dataset, 20% as test dataset and 30% as validation test. In contrast to [13], [14] they were extracted dataset from *Ask.fm* (question-answer based) using *GNU Wget* software and all of these data in *Dutch* language. By doing some refinement of *non-Dutch* data, the final posts are about 85,463. Other than that, [15] collected about 316,500 data from *Instagram* including images and comments which retrieved from 25,000 users. Besides that, [16] used *Twitter* as a data source. 1,762 tweets used as sample, which collected on *August* 2011.

Twitter dataset may easier to extracted compared to other mediums such as *Facebook*, *Instagram* and *YouTube*. Even though statisticbrain.com aforementioned stated that cyberbullying occurred most in *Facebook* but only data from public profiles could be extracted easily such as *Twitter* that the data is publicly available.

978-1-5386-7541-0/18/$31.00

B. Feature Used in Cyberbullying Detection

Before we going in-depth, we firstly categorize features used in cyberbullying detection studied based on [17]. There are four main categories; content, sentiment, user and network-based features.

Based on [12], authors mentioned three types of features have been used; profanity, negativity and subtlety. All of these features classified as content-based features. Three groups of topics were classified during annotation; intelligence, race & culture and sexuality.

In contrast, [13] and [14] were used two types of features which are content-based feature (BoW) and sentiment-based feature (polarity). Both studied stated if using single feature in order to detect cyberbullying was not enough because by integrating both features, result F-score shows high percentage instead of using features separately.

Besides, [15] mentioned a few features were used; cyberaggresion, profanity, network graph, image, and linguistic. Network graph involved number of likes, number of comments, number of followers and number of following. All of these features managed into a term: media session. According [15], fundamental discovery was mentioned by authors where researchers could not be depends only on profanity feature in order to enhance accuracy in cyberbullying detection. By analyzed network graph, media sessions consist of cyberbullying have low number of likes even though owners of media session have higher number of followers in Instagram account. Authors used Linguistic Inquiry and Word Count (LIWC) to extract linguistic features that is cyberbullying words. For image features, when a picture appear image like drug, then that image will be related to cyberbullying instead of picture contain image like scenery, book, etc. By combining three types of features; text, image and network graph, the authors concluded that text-based features could increase performance of cyberbullying detection instead of non-text based feature after implementing classifiers.

BoW features, Latent Semantic features and bullying features were used by [16]. The authors combined all three types of features as final representation called Embedded Bag-of-Word (EBoW). Precision, recall and F-score were higher when used EBoW instead of BoW, semantic BoW (sBoW), Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

C. Classification Used in Cyberbullying Detection

Based on [12], authors implemented binary classifier; Naïve Bayes, Rule-based JRiP, Tree-based J48 and Support Vector Machine (SVM). Two experiments were set up where first experiment used binary classifier to train three labels dataset (intelligence, culture & race and sexuality). Second experiment was integrated three datasets into a single dataset and trained using multiclass. Result shows that binary classifier with three label were better in terms of accuracy instead of using multiclass classifiers with one dataset. Accuracy of rule-based JRip was better than other binary classifiers.

In other studied by [13] and [14], they mentioned binary classifier (SVM) used as classification algorithm. By integrated BoW and polarity, result for F-scores was better than using single feature in SVM classification.

On the other hand, [15] implemented linear SVM, logistic regression, decision tree and *AdaBoost* as classifiers. However, only linear SVM gave a better result instead of decision tree and *AdaBoost*. Result of precision and recall for both linear SVM and logistic regression were quite similar for cyberaggresion detection.

Linear SVM was also used by [16] to learn featured. EBoW model showed better performance in terms of precision and recall compared to BoW, sBow, LSA and LDA when implemented linear SVM as classifier.

In summary, all of the studies used SVM as classification algorithm. Frequent use SVM by researchers shows that SVM is popular among others classifiers in supervised learning approach. SVM is suitable for high-skew text classification such as to detect cyberbullying using content-based features [18]. Any circumstances such as missing data, type of features and computer performance, SVM still outperform other classifiers [17]. Table 1 shows the summarization of data source, features used and classification in cyberbullying detection for each research works as discussed.

TABLE I. SUMMARIZATION OF STUDY IN CYBERBULLYING DETECTION

Study	Data Source	Feature	Classification
[12]	*YouTube* and *Formspring.me*	• Content-based feature (Profane, Negativity, Subtlety)	• SVM • Naïve Bayes • JRip • J48
[13]	*Ask.fm*	• Content-based feature (BoW) • Sentiment-based feature (Polarity)	SVM
[14]	*Ask.fm*	• Content-based feature (BoW) • Sentiment-based feature (Polarity)	SVM
[15]	*Instagram*	• Content-based feature (Profanity, Linguistic, Image, Cyberaggression) • Network-based feature (Network graph, Comments)	• Linear SVM • Logistic regression • Decision tree • *AdaBoost*
[16]	*Twitter* (http://research.cs.wisc.edu/bullying/data.html)	• Content-based feature (BoW, Bullying) • Latent semantic feature	Linear SVM

III. CHALLENGE IN CYBERBULLYING DETECTION

A. Language Challenge

In fact, the study in cyberbullying field is still immature in context of research world. For an example of a sentence such as "*The picture that you have sent so annoyed me and I do not want to contact with you anymore!*" is not easy to classify as cyberbullying without analyzing from a logic factor although that example show negative sentiment [17]. Every now and

then, positive message verse might be express with the intention of express sarcasm. Even so to detect cyberbullying is not easy by cause of nature of bullying which very subjective and subtle. In addition, with the modern world today, technology is expanding rapidly and likewise with language applied nowadays. In this regard, language used by adolescents change quickly and it will affected keywords implemented as feature in cyberbullying detection. Thus, supplementary factors may be required to prove such message labeled as cyberbullying.

B. Dataset Challenge

Another challenge in cyberbullying detection is dataset. To extract data from social media is not an easy task since it related to privacy information and social media sites do not reveal data openly. Consequently, this may cause lack information such as list of friends can be retrieve. In addition, annotation or data labeling is one tough task because it requires intervention from experts to label the corpus as studied by [19]. If there were potential researchers who can share the dataset that they have used, it would be a significant contribution to the world of research.

C. Data Representation Challenge

Most researchers only conduct research related to bullying words in telecommunication. Nevertheless, extracting content-based features have their own challenge. If an account of users does not provide information such as gender or age, performance in cyberbullying detection may deteriorate. But at the same time, [20] analyzing language utilized by users in order to determine range of age. It may take some time to identify word used in corpus that related to age. As an example, word 'study' may correlate to users in range 13 years – 18 years. In summary, to establish proper cyberbullying detection system or application is not easy since it involves human behavior and cyberbullying nature which difficult to interpret in context of cyberbullying.

IV. CONCLUSION

With the rapid technological growth, it is easier for users to widen their human network especially via social media. Conversely, if users abuse social media to commit cyberbullying, they can be categorized as barbaric fellow human being.

Mostly, researchers worked on identifying bullying keywords within the corpus using text classification in Natural Language Processing (NLP) and machine learning approaches. Hopefully, research on cyberbullying may be able to implement deep learning since it can work properly within text classification as studied by [21] for spam detection. In the future, research regards to cyberbullying may be able to collaborate with other field such as psychologist and sociologist to enhance the cyberbullying detection.

REFERENCES

[1] N. M. Zainudin, K. H. Zainal, N. A. Hasbullah, N. A. Wahab, and S. Ramli, "A review on cyberbullying in Malaysia from digital forensic perspective," *ICICTM 2016 - Proc. 1st Int. Conf. Inf. Commun. Technol.*, no. May, pp. 246–250, 2017.

[2] A. Saravanaraj, J. I. Sheebaassistant, S. Pradeep, and D. Dean, "Automatic Detection of Cyberbullying From Twitter," *IRACST - International J. Comput. Sci. Inf. Technol. Secur.*, vol. 6, no. 6, pp. 2249–9555, 2016.

[3] S. Hinduja and J. W. Patchin, "Cyberbullying: Identification, Prevention, & Response," *Cyberbullying Res. Cent.*, no. October, pp. 1–9, 2018.

[4] N. Willard, "Educator ' s Guide to Cyberbullying , Cyberthreats & Sexting," *Online*, pp. 1–16, 2007.

[5] Q. Li, "Bullying in the new playground: Research into cyberbullying and cyber victimisation," *Australas. J. Educ. Technol.*, vol. 23, no. 4, p. 435, 2007.

[6] N. E. Willard, "Overview of Cyberbullying and Cyberthreats," in *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*, Illinois: Reseach Press, 2007, pp. 5–16.

[7] A. Tooley, "Interview with Dr. Jonathan B. Singer on Cyberbullying," *Online MSW Programs*, 2018. [Online]. Available: https://www.onlinemswprograms.com/in-focus/interview-with-dr-jonathan-singer-on-cyberbullying.html. [Accessed: 28-May-2018].

[8] B. E. Palladino *et al.*, "Perceived severity of cyberbullying: Differences and similarities across four countries," *Front. Psychol.*, vol. 8, no. SEP, pp. 1–12, 2017.

[9] V. Dalla Pozza, M. Limited Anna Di Pietro, M. Limited Sophie Morel, M. Limited Emma Psaila, and M. Limited, "Cyberbulling among young people," 2016.

[10] J. Xu, K. Jun, X. Zhu, and A. Bellmore, "Learning from Bullying Traces in Social Media," *Proc. 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, pp. 656–666, 2012.

[11] S. B. R. Institute, "Cyberbullying/ Bullying Statistics," *Statistic Brain Research Institute*, 2018. [Online]. Available: https://www.statisticbrain.com/cyber-bullying-statistics/.

[12] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Commonsense Reasoning for Detection, Prevention and Mitigation of Cyberbullying," vol. 1, no. 212, 2012.

[13] C. Van Hee *et al.*, "Automatic detection and prevention of cyberbullying," *Int. Conf. Hum. Soc. Anal. (HUSO 2015)*, no. c, pp. 13–18, 2015.

[14] C. Van Hee *et al.*, "Detection and Fine-Grained Classification of Cyberbullying Events," *Int. Conf. Recent Adv. Nat. Lang. Process.*, pp. 672–680, 2015.

[15] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of Cyberbullying Incidents on the Instagram Social Network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9471, pp. 49–66, 2015.

[16] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," *Proc. 17th Int. Conf. Distrib. Comput. Netw. - ICDCN '16*, pp. 1–6, 2016.

[17] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," *IEEE Trans. Affect. Comput.*, pp. 1–25, 2017.

[18] B. Desmet and V. Hoste, "Recognising suicidal messages in Dutch social media," *Proc. Ninth Int. Conf. Lang. Resour. Eval.*, pp. 830–835, 2014.

[19] E. Raisi and B. Huang, "Cyberbullying Identification Using Participant-Vocabulary Consistency," pp. 46–50, 2016.

[20] H. A. Schwartz *et al.*, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PLoS One*, vol. 8, no. 9, 2013.

[21] G. Jain and B. Agarwal, "An Overview of RNN and CNN Techniques for Spam Detection in Social Media," vol. 6, no. 10, pp. 126–132, 2016.