Cyberbullying detection using web content mining*

Ana Kovačević*

Abstract —**Cyberbullying has become intensive field of research, due to its major impact on society. Most researchers analyze causes and consequences of cyberbullying, however, only few try to improve software to reduce or stop cyberbullying, and make Internet a safer place. In this article, current review of efforts in cyberbullying detection using web content mining techniques is presented.**

Keywords —***cyberbullying, information security, text mining, web mining.***

I. Introduction

Bullying is not a new phenomenon, since it has existed from ancient times. But, it has acquired a novel dimension with the rise of a new environment – the Internet, and developed a new form known as cyberbullying. According to the definition of the National Crime Prevention Council, cyberbullying is the use of the Internet, cell phones or other technologies to send or post a text or images intended to hurt or embarrass another person [1]. Cyberbullying can be carried out through several technology platforms, such as chat rooms, emails, photo sharing websites, blogs, forums, social networking cites, cell phones, online games and voice mail. Bullying in the cyber environment is much crueler and more dangerous than the "traditional" forms of bullying, which take place in the real world [2]. Reasons for that are primarily aspects of the web: the persistence, the ability to search and copy, as well as invisible audiences [3]. Because of the web persistence, a victim cannot hide anywhere, since the audience is not confined to a room, school yard or street, but presents a large online community. According to Dempsey et al. [4], two basic characteristics of cyberspace are dominant for cyberbullying: anonymity in cyberspace and better control of social interaction in the cyber world. Abusers can choose when they want to harass the victim, how (through which medium), and whether they wish to bully him/her in front of an audience.

Cyberbullying makes a major impact on society; consequently it has become intensive field of research. Although many researchers analyze causes and consequences of cyberbullying, only few suggest possible solutions for the prevention that include software systems beyond the usual ones based on key words [5]-[11].

This article reviews concepts of possible proposals how Internet may become a safer environment by using web content mining techniques for detecting and tracking cyberbullying. If the problem of cyberbullying comes to a solution or at least becomes minimized, social interaction would be much safer for many users on the web, especially for the most vulnerable ones, like teenagers.

II. Web Content Mining

Today the Internet is a very important part of every day's life, and a lot of data is generated. Discovering information from huge amount of data manually, may be extremely complicated, frequently impossible. Researchers in cyber security face increasing amounts of information and it is evident that more powerful tools are needed to handle that [12]. Web mining tools can help in discovering new, previously hidden information from a huge amount of data [13]. Web mining is the application of data mining on content, structure and usage of web recourse to discover knowledge from data on web. Usually, according to the analyzed data, web mining is classified in the following categories [14]:

- Web content mining as a process of discovering useful information from content of web document.
- Web structure mining, as a process of discovering structural information of World Wide Web.
- Web usage mining as the use of data mining techniques to find user's patterns of weblogs, for improving web applications.

Web content mining analyses unstructured data (text, image, or video) or structural records (e.g., lists and tables). Most research is done in the domain of text, using text mining techniques for extracting relevant facts from web documents or finding relevant web documents. Text mining is considered as sub category of web content mining that does not use web structure [15]. Text mining discovers hidden information using methods that are, on the one hand, able to analyze many words and structures in natural language and, on the other hand, can cope with vagueness, uncertainty and often fuzziness within the text documents [16].

Conventional analyses and data mining techniques used for traditional Web content are not suitable for the Web 2.0. The main reason for that is the fact that the Web 2.0

* Acknowledgment: Paper is the result of the research within the project 47017, that are financed by Serbian Ministry of Education and Science

* Ana Kovačević is with the Faculty of Security Studies, University of Belgrade, Gospodara Vučića 50, 11000 Beograd, Serbia (phone: 381.11.6451.963; e-mail: kana@rcub.bg.ac.rs)

contents are generated by the users, who opt for informal textual communication, e.g., a very free use of language, and therefore constantly incorporate new communication elements which are generally context dependent. This kind of language can also be found in chats, SMS, e-mails etc.

Text analysis of Web 2.0 is further complicated by the fact that language used in online chats and social networking sites is the jargon, often with syntax errors and a common use of emotions. There are "friendly insults", which are not considered as cyberbullying, but as a style of communication. Sarcasm additionally makes it difficult to detect text mining. Moreover, a common phenomenon is that teenagers exchange sarcastic comments without the intention to hurt one another, and these cases should not be misinterpreted as cyberbullying.

III. Text Mining for Detection of Cyberbullying

There are several research teams working on the detection of cyber abuse using text mining techniques. Their research is related to applications in various technology platforms (e.g., social networking cites, chat rooms, forums), and some of the basic approaches are presented here.

A. Detection of cyberbullying in messages

Yin et al., conducted experiments on three different data sets (My Space, Slashdot and Congregate) provided by Content analysis for Web 2.0 (CAW 2.0 http://caw2.barcelonamedia.org/) in order to detect harassment [5]. For harassment detection they used content, sentiment, and contextual features of the documents to train a support vector machine (SVM) classifier for a corpus of online posts. Various methods were used to develop the attributes of the entrance to the classifier, such as: standard text mining techniques based on weights of term (in this case - words), rule-based systems for detection of feelings and context analysis. The obtained results demonstrated that the use of the combined model, which besides text mining included methods for adding context and detection of feelings, improved the detection of cyberbullying.

B. Detection of cyberbullying in comments form YouTube video clips

Researchers at the MIT (web.mit.edu) detected the cyberbullying in comments from YouTube videos [6]. They assumed that a comment could be a form of bullying if one of the sensitive features were contained, with a negative tone. Sensitive characteristics are one of the following: physical appearance, sexuality, race / culture and intelligence.

The comments were firstly manually categorized by assigning labels from sensitive characteristics. They made two experiments. Firstly, they had analyzed whether the comments corresponded to anything related to the sensitive issues like sexuality, race / culture, intelligence or physical characteristics. Secondly, they determined the theme. Supervised learning methods used in the experiment were: Repeated Incremental Pruning to Produce Error Reduction (JRip) [17] J48 [18] and SVM for classification [19].

The authors presented that binary classifier trained for individual labels can outperform the detection of textual cyberbullying compared to multiclass classifier.

C. Program BullyTracer

The BullyTracer program is designed to detect different types of cyber harassment in a chat on the MySpace website [7]. A rule-based algorithm was used to detect cyberbullying. BullyTracer uses terms (words) from the selected dictionary, divided in three categories: insulting words, vulgar language and pronoun “YOU”. Authors gave examples of two posts:

- Post1: “you are so dumb you make me want to kill someone”
- Post2: “that test was so dumb it was killing me.”

Although these two posts have many words in common *Post1* should be consider as cyberbullying, and *Post2* should not. The authors selected these categories since they noticed there was a significant correlation between the presence of these words and the occurrence of cyberbullying. BullyTracer program was used to examine considered posts and search through them for words from the above mentioned categories (e.g., insulting words or swearing). Posts that contained such wording were marked by the program.

D. Detection of cyberbullying on Twitter

A framework for the detection of cyberbullying on Twitter was created by Sanchez & Kumar [8]. Text that was used in messages (tweets, twitter message) requires intensive pre-processing prior to classification, including identification of syntax errors, emotions, and use of slang. The idea was to classify emotions (especially negative) contained in a message using a sentiment analysis and opinion mining, and then to visualize the changes in the message over time.

The messages were classified using NaiveBayes algorithm as negative or positive, with respect to some frequently used words. Bag-of-words model was used in the classification. The aim of the authors was to identify the victims, followers and predators. Following the identification of cyberbullying visualization was applied. By using dynamic visualization, cyberbullying was tracked down and illustrated over time.

E. Detection of cyberbullying on informal web site

Problems of cyberbullying in Japanese society, particularly on unofficial scholars Web sites, were analyzed by Ptaszynski et al. [9]. The machine learning method developed to handle cyberbullying activities consists of several stages: creation of lexicon comprising vulgar words (in Japanese), slanderous information detection module, the information ranking in accordance with/(in regard to) the level of their harmfulness, and visualization of the harmful information. After the message is classified as cyberbullying or not, using SVM algorithm, it is also important to determine how harmful a certain entry is.

F. Detection of cyberbullying in comments form YouTube video clips

Dadvar et al. [10] detected cyberbullying in comments from YouTube video. They used combination of content-based, cyberbullying-specific and user based features. They have shown that using user context (user's comments history and user characteristics) improves cyberbullying detection accuracy. For training they used SVM binary classifier .

G. Detection of cyberbullying in social networks

Nahar et al. [11] proposed semi-supervised approach for detection in social networks, by devised new framework for automatic detection of cyberbullying for streaming data with insufficient labeling. They conducted experiments on three different data sets (My Space, Slashdot and Congregate) provided by Content analysis for Web 2.0 (CAW 2.0 http://caw2.barcelonamedia.org/). The enriched features sets were generated based of user context, linguistic knowledge and baseline keywords. They proposed fuzzy SVM algorithm for cyberbullying detection.

H. Discussion

Table 1. Some characteristics of proposed solutions

Ref	*Data sets*	*Characteristics*	*Classifier*
[5]	CAW2.0* (My Space Slashdot Congregate)	Rule based detection of feelings Context analysis English language	SVM
[6]	YouTube comments	Sensitive issues (sexuality, race/culture, intelligence or physical characteristics). English language	SVM J48
[7]	Chat on the MySpace	Rule based algorithm use terms (insulting words, vulgar language, pronoun "YOU") English language	--
[8]	Twitter	Classify emotions Sentiment analysis and opinion mining Visualization English language	NaiveBayes
[9]	Unofficial scholars Web sites	Creating lexicon of comprising vulgar words (in Japanese), Information ranking Japanese language	SVM
[10]	YouTube comments	Content based, cyberbullying-specific and user based features. English language	SVM
[11]	CAW2.0* My Space Slashdot Congregate	User context, linguistic knowledge and baseline keywords. English language	Fuzzy SVM

Legend: * CAW2.0 was publicly available data set for (http://caw2.barcelonamedia.org/)

As it is shown in Table 1. the vast majority of researchers use mostly SVM algorithm for binary classification (if it is cyberbullying or not). Few authors used visualization as a powerful technique for an easier detection of cyberbullying [8],[9], and that method may be especially significant for moderators of social networking cites. All classifier are binary, (if it is cyberbullying or not), except in [6] where is used binary and multiclassifier.

Researchers using different data sets, so it is hardly to compare results. Existing studies of cyberbullying detection are manly use supervised learning approaches and assuming that data are pre-labeled, except in [11] where semi-supervised learning approach are used.

This is quite new field of research. One of the first review of using text mining techniques in detecting cyberbullies and cyber predators is analyzed in [20], and they only presented three research article.

IV. Conclusion and Future Work

Cyberbullying is a problem that occurs on the Internet, and it is of great importance to make Internet a safer environment especially for the most vulnerable population, such as teenagers. Since the identifying of cyberbullying is difficult, results indicate that it is possible to detect cyberbullying using web content mining techniques. Although a satisfactory level of accuracy has not been reached, the results are promising. Once, when the level of accuracy is achieved it may be implemented in software for social networks. Also, these efforts may be a useful tool for learning about patterns in cyberbullying or making better dictionaries of offensive words.

Lieberman et al. compares cyberbullying with spam at the beginning of using e-mail, where the spam threatened to hamper access and completely disable the use of electronic mail [21]. Thanks to the development of efficient algorithms for detecting spam, now spam exists in some negligible quantity.

In future studies, user's and context characteristics will be more included in automatic detection of cyberbullying, as well as (semi)-automatic labeling data prior to classifying. Also, new models that include sarcasm and implicit harassment should be developed. Developing new frameworks for cyberbullying detection for non-English language should be also expected. Visualization may be useful in easier notification about the presence of bullies.

After detecting cyberbullying, appropriate action could be taken, such as preventing further abuse of the victim, slowing the spread of potentially offensive messages, and providing additional educational materials to assist victims and facilitate the problem.

Acknowledgment

Paper is the result of the research within the project 47017, that are financed by Serbian Ministry of Education and Science.

References

[1] NCPC (2006) Cyberbullying, available at http://www.ncpc.org/cyberbullying

[2] D. Boyd, Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life, MIT Press., 2007.

[3] A. Kovacevic, D. Nikolic, “Automatic detection of cyberbullying to make Internet a safer environment”. Handbook of Research on Digital Crime”, in Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance (pp. 1-675). editores Cruz-Cunha, M. M., & Portela, I. M., Hershey, PA: IGI Global, 2015

[4] A. Dempsey, M. Sulkowski, J. Dempsey, E. Storch, Has Cyber Technology Produced a New Group of Peer Aggressor, Cyberpsychology, behavior, and social networking, 2011, 14(5), 297-302.

[5] D. Yin, Z. Xue, L. Hong, B. Davison, A. Kontostathis, L. Edwards. Detection of Harassment onWeb 2.0. In CAW 2.0 ‘09: Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain 2009.

[6] K. Dinakar, R Reichart, H. Lieberman. Modeling the detection of textual cyberbullying, International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain 2011

[7] J. Bayzick, A. Kontostathis, L. Edwards. Detecting the Presence of Cyberbullying Using Computer Software, WebSci ’11, June 14-17, 2011, Koblenz, Germany 2011.

[8] H. Sanchez, S. Kumar, Twitter Bullying Detection, 2011. http://users.soe.ucsc.edu/~shreyask/ism245-rpt.pdf 2014.07.17.

[9] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, K. Machine Learning and Affect Analysis Against Cyberbullying, Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK.

[10] M. Dadvar, D. Trieschnigg, R. Ordelman, & F. de Jong, Improving cyberbullying detection with user context. In *Advances in Information Retrieval* (pp. 693-696). Springer Berlin Heidelberg, 2013.

[11] V. Nahar, S. Al-Maskari, X. Li, and C. Pang. "Semi-supervised Learning for Cyberbullying Detection in Social Networks." In Databases Theory and Applications, pp. 160-171. Springer International Publishing, 2014.

[12] A. Kovačević „Primena web mininga i vizualizacije u otkrivanju sajber nasilja” u Kordić, B., Kovačević, A. i Banović, B. (ur.), *Reagovanje na bezbednosne rizike u obrazovno-vaspitnim ustanovama*, Fakultet bezbednosti:Beograd, 2012, in Serbian.

[13] A. Kovacevic, A. Kovacevic, Vizuelizacija web mininga nad edukativnim objektima, FON, 2010, Doktorska disertacija, in Serbian.

[14] J. Srivastava, P. Desikan, V. Kumar, Web Mining: Accomplishments and Future Directions, in Proceedings US Nat’l Science Foundation Workshop on Next-Generation Data Mining (NGDM), National Science Foundation 2002.

[15] J. Kolari, A. Joshi, Web mining: research and practice, IEEE Computing in Science and Engeneering, July/Aug 2004.

[16] A. Hotho, A. Nürnberger, G. Paaß. A Brief Survey of Text Mining. Forum American Bar Association, 2005, 20(1), 19-62.

[17] W.W. Cohen,W.W., & Y. Singer. A simple, fast, and effective rule learner. In Proceedings of the Sixteenth National Conference on Artificial Intelligence, 1999.

[18] R. Quinlan. C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA 1993.

[19] C. Cortes, V. Vapnik "Support-Vector Networks", Machine Learning, 20, 1995.

[20] A. Kontostathis, L. Edwards, and A. Leatherman. "Text mining and cybercrime." Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK (2010).

[21] H. Lieberman, K. Dinakar & B. Jones (2011). Let's Gang Up on Cyberbullying. Computer, 2011, 44(9), 93-96.