Optimized Twitter Cyberbullying Detection based on Deep Learning

Monirah A. Al-Ajlan
Information Systems Department, CCIS
King Saud University
Riyadh, Saudi Arabia
maalajlan@ksu.edu.sa

Mourad Ykhlef
Information Systems Department, CCIS
King Saud University
Riyadh, Saudi Arabia
ykhlef@ksu.edu.sa

Abstract— **Cyberbullying is a crime in which a perpetrator targets a person with online harassment and hate. Many cyberbullying detection approaches have been introduced, but they were largely based on textual and user features. Most of the research found in the literature aimed to improve detection by introducing new features. Although, as the number of features increases, the feature extraction and selection phases become harder. In addition, another drawback of such improvements is that some features—for example, user age—can be easily fabricated. In this paper, we propose optimised Twitter cyberbullying detection based on deep learning (OCDD), a novel approach to address the above challenges. Unlike prior work in this field, OCDD does not extract features from tweets and feed them to a classifier; rather, it represents a tweet as a set of word vectors. In this way, the semantics of words is preserved, and the feature extraction and selection phases can be eliminated. As for the classification phase, deep learning will be used, along with a metaheuristic optimisation algorithm for parameter tuning.**

Keywords- cyberbullying; detection; deep learning; convolutional neural network; optimization algorithm

INTRODUCTION

Technology is dominating our lives today; we rely upon technology to carry out most of our daily activities. Communication is no exception, as technology has changed how people interact in a very broad manner and has given communication a new dimension. As promising as it sounds, this huge shift from traditional to digital world has come with a pricy cost. The anonymous nature of social networks, where users use nick names rather than their real ones making their actions very hard to trace has resulted in a growing number of online crimes like cyberbullying. Cyberbullying is one of the top ethical issues found on the Internet, and the percentage of people, especially teenagers who were victims of cyberbullying is alarming. Cyberbullying is defined as any violent, intentional action conducted by individuals or groups, using online channels repeatedly against a victim who does not have the potential to react[1]. Many studies have addressed cyberbullying with aim of assessing its prevalence, and results showed that cyberbullying is a common issue facing today's generation and that the number of victims is rising [2][3]. In order to help in controlling cyberbullying and limiting its prevalence, many cyberbullying detection mechanisms have been introduced.

The work of researchers in the field of cyberbullying detection has evolved however, many problems have not been addressed yet. First, the language used changes over time, so there is no static list of bad words -words that would be considered as used in the effort to cyberbully- that can be considered. Second, the semantic of words has not been considered. Detection was carried out by extracting features from text and building a classifier accordingly. Third, cyberbullying detection achieved by the use of a classifier was always improved by adding features, therefore, the feature extraction and selection phases have become harder to complete and more time consuming. To address these problems, the proposed approach eliminates the feature extraction and selection phases and replace them by the use of word vectors which will be fed to a convolutional neural network (CNN) for classification. A method where semantics are preserved and classification is accomplished without features. CNN has many parameters and the choice of their values is critical to the achieved result. Therefore, metaheuristic optimization algorithm is incorporated to find the optimal or near optimal values. It is expected to overcome current limitations and improve the detection accuracy.

The proposed approach (OCDD) answers the following research questions:

1. Does CNN give better classification results?
2. Does metaheuristic optimization algorithm find optimal values for the parameters of CNN?

The reminder of the paper is organized as follows. Section 2 covers background about cyberbullying and how it is detected. In section 3, the literature review about cyberbullying detection in social media is discussed and

978-1-5386-4110-1/18/$31.00 ©2018 IEEE

compared. Then, section 4 illustrates our proposed approach (OCDD) and describes its phases. Finally, section 5 is the conclusion of this paper.

BACKGROUND

A. Cyberbullying

A large and growing body of literature has investigated that Cyberbullying does not have one specific definition, rather a large and growing body of literature has investigated it from different perspectives and a variety of definitions have been suggested. According to Kowalski [4], cyberbullying is the online misbehavior carried out and facilitated by means of Information and Communication Technologies (ICT) including text messages and social network sites. Another definition found in [5], is that cyberbullying 'an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself'.

With the advances of technology and its wide spread reach, many ethical issues like cyberbullying has proliferated. Social networks have been used as a platform for teens wanting to harm others [6]. Figure 1 shows the result of a conducted by the Cyberbullying Research Group on a sample of 5700 US school students aged between 12 to 17 as illustrated in Fig. 1 showed that more that 33% of students have experienced some form of cyberbullying in their life [3]. Another study conducted by the National Center of Social Science (NatCen) reported similar result [7]. They have focused their work on two sub categories of US school students, those reported as frequently absent from school and those who opted for home school programs. The reason behind their selection lies in their research objective which was to measure whether cyberbullying lead victims to be absent from school or even to leave school in more serious scenarios. This work yielded that 17.5% of students frequently get absent due to the negative consequences it leaves. Moreover, students' parents considered cyberbullying as the fifth major reason for choosing home schooling programs for their kids. These fundemental findings emphasize the fact that cyberbullying is a critical concern affecting large population.

On a local scale, in Saudi Arabia a study conducted in 2013 by the National Family Safety Program (under the supervision of the Ministry of Labor and Social Development) on a sample size of 15264 high school students showed that 25% experienced cyberbullying during their life [8]. As a result, cyberbullying negative effects have been recognized in Saudi Arabia and a special campaign named the National Project for Cyberbullying Control was launched in 2014 by the National Family Safety Program to reduce cyberbullying negative effects and raise the society awareness of cyberbullying toward developing students' self-esteem.

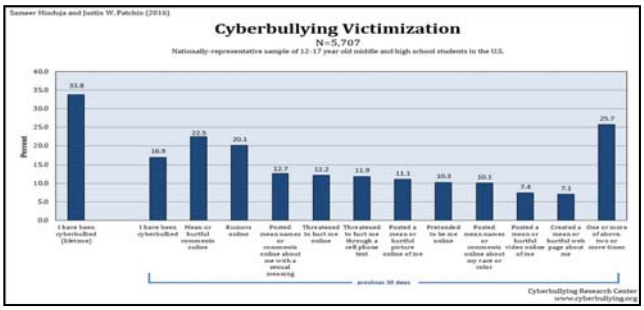

Figure 1. Cyberbullying victimization chart. [2]

B. Metaheuristics Optimization

Algorithms that find solutions but do not guarantee that they are the best solutions are called heuristic algorithms. They accomplish this by replacing original optimization goal with an alternative goal, one that is easier to reach. Metaheuristic optimization is widely used to find the best (or near best) solution for a given problem. Problem can be defined as any task that needs to be solved given a space of solutions and a search mechanism. The necessity for optimization has arisen when solution spaces became huge and classical search algorithms failed to function [9]. Since 1970s, many algorithms named evolutionary algorithms have been proposed to solve optimization problems by mimicking natural phenomena. Examples of such algorithms are: Firefly algorithm (FA), Bees algorithm (BA) and Ant Colony optimization algorithm [11]. Naturally inspired algorithms or evolutionary algorithms have been widely adopted in data mining, they showed great improvements when employed in feature selection and classification, two areas of data mining usually suffer from high dimensionality and large solution space [10].

C. Convolutional Nueral Network (CNN)

CNN is a special case of neural network which is capable of understanding the inner structure of data like the structure of an image. The architecture of a CNN is mainly a neural network with many pooling and convolutional layers. The convolutional layer is responsible for taking a subset of data (pixels in case of an image) and perform some mathematical calculations on the subset of data. The result of the calculations is another image which is then passed to a pooling layer that reduces the size of image by combining adjacent pixels [11].

CNN operates on text as well as in images and many researchers employed CNN for various text mining tasks [11] [12] [13]. The difference is that in case of text, vector of pixels will be substituted by a vector of word embedding and the rest of CNN structure will be the same.

LITERATURE REVIEW

Bullying detection falls under the umbrella of event detection; a broader research area concerned with identifying extraordinary events given a mass data and noisy content [13]. Event detection considers events that are characterized as being a thing happening at a specific point of time [14] and can be expressed via traditional media channels [13]. In [15], researchers addressed the issue of identifying critical events in Twitter when they are overwhelmed by irrelevant tweets, their work focused on combining textual analysis and social aspects of Twitter. They proposed MABED, a novel event detection methodology that analyzes tweets through factors like mentions and URLs to statistically detect events A similar study was conducted [14], where they aimed at classifying tweets based on whether they describe real-world events or not. Another research that addresses event detection is found in [16], where they proposed a system called (TEDAS), which is capable of detecting events, analyzing tweets patterns and evaluating the importance of the detected event.

Cyberbullying detection -a sub field of event detection- has a rapidly growing literature, even though researches addressing bullying are traced back to early 2010. Among the first to tackle bullying in social media is [17], where a framework was built to incorporate Twitter streaming API for collecting tweets and then classifying them according to the content. Their work combined the essence of sentiment analysis and bullying detection. As a first phase, tweets are classified as being positive or negative and then they are further classified as positive containing bullying content, positive without bullying content, negative containing bullying content, and negative without bullying content. For the sake of classification, Naïve Bayes was implemented and resulted in a relatively high accuracy (70%). Another research [18], presented a prototype system to be used by organization members to monitor social network sites and detect bullying incidents. The approach followed relied on recording and storing bullying words and storing them in a database and then incorporate Twitter API to capture tweets and compare their content to the bullying material recorded earlier. If a bullying incident is detected, an email will be delivered to the police station concerned with electronic crimes. Beside the promising innovative idea in their work, this prototype system has not been implemented yet. A broader perspective on bullying detection was conducted in [19], where a survey comparing the accuracy of many classification algorithms (SVM, J48 and Naïve Bayes) revealed that SVM with linear kernel resulted in the best accuracy reaches (81.6%). This work is deemed to be a reference for cyberbullying detection as they presented a complete framework including data collection, preprocessing, detection and classification. The authors of [19] compared SVM, KNN and Naïve Bayes and SVM showed best accuracy.

Features found to be influencing cyberbullying detection have been explored in several studies. One of the early attempts to evaluate the quality of features is found in [25], where they assumed that cyberbullying detection is largely based upon language features. Their work considered three different language features: 1) the number of bad words (NUM) 2) the density of bad word (NORM) and 3) the overall badness of a tweet (SUM). It was reported that NORM feature produced the highest accuracy (81.7%) when run on different classification algorithms. Even though the reported accuracy was not high; this result (considering textual features only) emphasized that language is a major area of interest within the field of cyberbullying detection and has led to a proliferation of studies that are concerned with language processing and understanding.

Despite the textual features success in detecting cyberbullying, researchers investigated other kinds of features and evaluated their contribution. Many researches that addressed cyberbullying along with the features they considered are presented in Table 1.

TABLE 1 CYBERBULLING RESEARCH SUMMARY

Paper	Data Mining Task	Algorithm Used	Features
[20]	Classification	SVM, J48, Naïve Bayes	- Text - Profile - User graph
[21]	Classification	SVM.	- Text - Bag of words - Author
[22]	Classification	Lib SVM	- LDA /TFIDF - List of bad words - pronouns
[23]	Classification	J48, Naive Bayes, SMO, Bagging and Dagging.	- Social network - Text - Part of speech
[24]	Classification	SVM	- Text - Sentiment
[25]	Classification	C 4.5	- Text - List of bad words
[26]	Classification	SVM	- User - Content - Cyberbullying based
[27]	Classification	SVM, Naive Bayes	- Content - Profile
[28]	Classification Clustering	Fuzzy SVM	- User - Location - Text - Media
[29]	Classification	Fuzzy logic	- Text - Word statistics
[30]	Classification	SVM, J 48	- Text - TFIDF

Overall, these studies highlighted the fact that bullying detection has not been investigated deeply and that it is relatively a new hot research area. In view of all that has

been mentioned so far, optimization was never used to enhance the classification process nor deep learning and its various forms.

PROPOSED APPROACH

The proposed approach aims to improve cyberbullying detection by enhancing prior work which did not attempt to incorporate semantics and required hard and long feature extraction. In order to fill this gap, the classification of bullying/nonbullying tweets will be enhanced using convolutional neural network. Fig. 2 shows the system architecture.

As a first step, data collection is performed using twitter4j API [31] and a java code was written to fetch 20,000 random tweets. Then, data cleaning step was done to remove noisy, irrelevant and duplicate tweets. These data were then split into training and testing sets. As for the training data, tweets annotation is required to label tweets into (bullying-nonbullying), and CrowdFlower [32] a trusted data science and machine learning platform was used for this task.

In order to capture the semantics and similarities between words, word embedding is used to model words. Vectors can be incorporated as features for many data applications like information retrieval [33] and question answering [34]. For the sake of word embedding, Global Vector (Glo Ve) was chosen because it has outperformed other models in word similarity and named entity recognition [25].

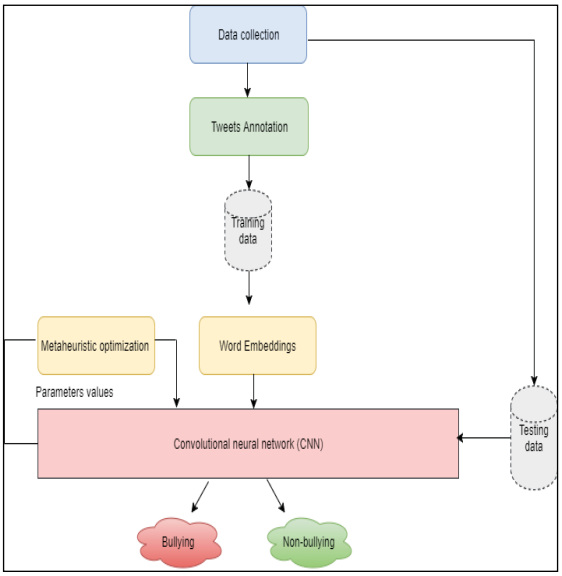

Figure 2. system architecture

The proposed approach uses a pre-trained (Glo Ve) which have two billion tweets.

After representing tweets as word vectors, they are fed to a convolutional neural network (CNN) which will learn through iterations, patterns of data. However, CNN has many parameters like the number of epochs, neurons per layer and number of layers (depth). In our experiments, it has been shown that changing the values of parameters changes the classification accuracy therefore, optimization algorithm is essential to find an optimal or near optimal values for the parameters. To address this issue, metaheuristic optimization algorithm was added.

Metaheuristic optimization algorithms that mimics the behavior of natural insects are based on the idea that a population must be specified. In our case, a set of CNNs with random values of parameters represent the population. In order to determine how population will be evolved, an objective function has to be specified which is the classification accuracy. Eventually, metaheuristic optimization algorithm will find the optimal or near optimal set of values that will be used for classification.

Cyberbullying detection is mainly a classification problem, therefore, evaluation is accomplished by recording three measures: accuracy, precision and recall (also called sensitivity). All experiments will be done on a single processor PC with 12 GB of RAM.

CONCLUSION AND FUTURE WORK

Social media has changed the way people communicate, it made everyone exposed to many forms of danger like cyberbullying. Even though many researchers addressed cyberbullying in social media, the classification of (bullying-nonbullying) was heavily dependent on the quality of features fed to the classifier.

In the proposed approach, we explored the current Twitter cyberbullying detection techniques, and proposed a new classification method based on deep learning. Our proposed approach (OCDD) was built using training data labelled by human intelligence service and then word embedding was generated for each word using (Glo Ve) technique. The resulted set of word embedding was later fed to convolutional neural network (CNN) algorithm for classification. CNN operates on many different layer types each having different parameters to be set. Manual attempts can be extremely hard and slow therefore, metaheuristic optimization algorithm is incorporated to find the optimal or near optimal values. OCDD advances the current state of cyberbullying detection by eliminating the hard task of feature extraction/selection and replacing it with word vectors which capture the semantic of words and CNN which classifies tweets in a more intelligent way than traditional classification algorithms. CNN showed great results when used with different text mining tasks; however, it has not been implemented in cyberbullying detection context.

As for future work, we would like to adapt the proposed approach to allow Arabic content. Arabic language has different structure and rules so comprehensive Arabic natural language processing should be incorporated.

References

[1] R. Shetgiri, "Bullying and Victimization Among Children", Advances in Pediatrics, vol. 60, no. 1, pp. 33-51, 2013.

[2] Brown, E. Clery and C. Ferguson, "Estimating the prevalence of young people absent from school due to bullying," National Center for Social Research, 2011.

[3] Van Royen, K. Poels, W. Daelemans and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," Telematics and Informatics, vol. 32, no. 1, pp. 89-97, 2015.

[4] R. Kowalski, G. Giumetti, A. Schroeder and H. Reese, "Cyber Bullying Among College Students: Evidence from Multiple Domains of College Life," Misbehavior Online in Higher Education, vol. 1st, pp. 293-321, 2017.

[5] R. SLONJE and P. SMITH, "Cyberbullying: Another main type of bullying?," Scandinavian Journal of Psychology, vol. 49, no. 2, pp. 147-154, 2008.

[6] R. Donegan, "Bullying and Cyberbullying: History, Statistics, Law, Prevention and Analysis," The Elon Journal of Undergraduate Research in Communications, vol. 3, no. 1, 2012.

[7] K. Van Royen, K. Poels, W. Daelemans and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," Telematics and Informatics, vol. 32, no. 1, pp. 89-97.

[8] "National Project for Cyberbullying Control," [Online]. Available: https://nfsp.org.sa/ar/community/projects/project3/Pages/default.aspx. [Accessed 20-1-2017].

[9] Z. Michalewicz, Genetic algorithms + data structures = Evolution Program, Berlin: Springer-Verlag, 1969.

[10] Z. Woo Geem, J. Hoon Kim and G. Loganathan, "A New Heuristic Optimization Algorithm: Harmony Search," Simulation, vol. 76, no. 2, pp. 60-68, 2001.

[11] Johnson, Rie, and Tong Zhang. "Effective use of word order for text categorization with convolutional neural networks." arXiv preprint arXiv:1412.1058 (2014).

[12] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." Advances in neural information processing systems. 2015.

[13] A. J. McMinn, Y. Moshfeghi and J. M. Jose, "Building a large-scale corpus for evaluating event detection on twitter," in 22nd ACM international conference on Information & Knowledge Management, 2013.

[14] H. Becker, M. Naaman and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter," Association for the Advancement of Artificial Intelligence, 2011.

[15] A. Guille and C. Favre, "Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach," Social Network Analysis and Mining, vol. 5, no. 1, 2015.

[16] R. LI, K. Hou Lei, R. Khadiwala and K. Chen-Chuan Chang, "TEDAS: a Twitter Based Event Detection and Analysis System," ieee 28th international conference on Data engineering (icde), pp. 1273-1276, 2012.

[17] H. Sanchez and S. Kumar, "Twitter Bullying Detection," ser. NSDI, p. 15, 2011.

[18] L. Choong Hon, and K. Varathan, "CYBERBULLYING DETECTION ON TWITTER", International Journal of Information Systems and Engineering, vol. 3, no. 1, pp. 36-47, 2015.

[19] R. Sugandhi, S. Chawla, A. Agrawal, H. Bhagat and A. Pande, "Methods for Detection of Cyberbullying: A Survey," 15th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 173-177, 2015.

[20] E. Abozinadah, A. Mbaziira And J. Jones, "Detection Of Abusive Accounts With Arabic Tweets," International Journal Of Knowledge Engineering-Iacsit, vol. 1, no. 2, pp. 113-119, 2015.

[21] S. Sood, E. Churchill And J. Antin, "Automatic Identification Of Personal Insults On Social News Sites," Journal Of The American Society For Information Science And Technology, vol. 63, no. 2, pp. 270-285, 2011.

[22] V. Nahar, X. Li And C. Pang, "An Effective Approach For Cyberbullying Detection," Communications In Information Science And Management Engineering, vol. 3, no. 5, pp. 238-247, 2013.

[23] Q. Huang, V. Singh And P. Atrey, "Cyber Bullying Detection Using Social And Textual Analysis," 3rd International Workshop on Socially-Aware Multimedia, pp. 3-6, 2014.

[24] R. Sugandhi, A. Pande, A. Agrawal And H. Bhagat, "Automatic Monitoring And Prevention Of Cyberbullying," International Journal Of Computer Applications, vol. 144, no. 8, pp. 17-19, 2016.

[25] K. Reynolds, A. Kontostathis And L. Edwards, "Using Machine Learning To Detect Cyberbullying," 10th International Conference on Machine learning and applications and workshops (ICMLA), vol. 2, pp. 241-244, 2011.

[26] M. Dadvar, D. Trieschnigg, R. Ordelman And F. De Jong, "Improving Cyberbullying Detection With User Context," European Conference on Information Retrieval, pp. 693-696, 2013.

[27] H. Hosseinmardi, S. Mattson, R. Ibn Rafiq, R. Han, Q. Lv And S. Mishra, "Detection Of Cyberbullying Incidents On The Instagram Social Network," Association For The Advancement Of Artificial Intelligence, 2015.

[28] V. Nahar, S. Al-Maskari, X. Li And C. Pang, "Semi-Supervised Learning For Cyberbullying Detection In Social Networks," Australasian Database Conference, pp. 160-171, 2014.

[29] B. Nandhini And J. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques," Procedia Computer Science, vol. 45, pp. 485-492, 2015.

[30] K. Dinakar, B. Jones, C. Havasi, H. Lieberman And R. Picard, "Common Sense Reasoning For Detection, Prevention, And Mitigation Of Cyberbullying," Acm Transactions On Interactive Intelligent Systems, vol. 2, no. 3, pp. 1-30, 2012.

[31] "Twitter4J - A Java library for the Twitter API", Twitter4j.org, 2018. [Online]. Available: http://twitter4j.org/en/index.html. [Accessed: 14- Jan- 2018].

[32] "Training data, machine learning and human-in-the-loop for A.I.", CrowdFlower, 2018. [Online]. Available: https://www.crowdflower.com/. [Accessed: 14- Jan- 2018].

[33] Tellex, Stefanie, et al. "Quantitative evaluation of passage retrieval algorithms for question answering." Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003.

[34] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

[35] Ariyaratne, M. K. A., and T. G. I. Fernando. "A Comparative Study on Nature Inspired Algorithms with Firefly Algorithm." International Journal of Engineering and Technology 4.10 (2014).