Computers in Human Behavior 63 (2016) 433–443

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network

Mohammed Ali Al-garadi*, Kasturi Dewi Varathan, Sri Devi Ravana

Department of Information System, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form
11 May 2016
Accepted 17 May 2016
Available online 31 May 2016

Keywords:
Online social networks
Cybercrime
Cyberbullying
Machine learning
Online communication
Twitter

ABSTRACT

The popularity of online social networks has created massive social communication among their users and this leads to a huge amount of user-generated communication data. In recent years, Cyberbullying has grown into a major problem with the growth of online communication and social media. Cyberbullying has been recognized recently as a serious national health issue among online social network users and developing an efficient detection model holds tremendous practical significance. In this paper, we have proposed set of unique features derived from Twitter; network, activity, user, and tweet content, based on these feature, we developed a supervised machine learning solution for detecting cyberbullying in the Twitter. An evaluation demonstrates that our developed detection model based on our proposed features, achieved results with an area under the receiver-operating characteristic curve of 0.943 and an f-measure of 0.936. These results indicate that the proposed model based on these features provides a feasible solution to detecting Cyberbullying in online communication environments. Finally, we compare result obtained using our proposed features with the result obtained from two baseline features. The comparison outcomes show the significance of the proposed features.

1. Introduction

Online social networking sites have become immensely popular in the last few years. Millions of users have used these websites as novel communication tools and as real-time, dynamic data sources where they can create their own profiles and communicate with other users regardless of geographical location and physical limitations. In this regard, these websites have become vital, ubiquitous communication platforms. The communication data from online social networks can provide us with novel insights into the construction of social networks and societies, which is previously thought to be impossible in terms of scale and extent. Moreover, these digital tools can transcend the boundaries of the physical world in studying human relationships and behaviors (Lauw, Shafer, Agrawal, & Ntoulas, 2010).

Cyber criminals have utilized social media as a new platform in committing different types of cybercrimes, such as phishing (Aggarwal, Rajadesingan, & Kumaraguru, 2012), spamming (Yardi, Romero, & Schoenebeck, 2009), spread of malware (Yang, Harkreader, Zhang, Shin, & Gu, 2012), and cyberbullying (Weir, Toolan, & Smeed, 2011). In particular, cyberbullying has emerged as a major problem along with the recent development of online communication and social media (O'Keeffe & Clarke-Pearson, 2011). Cyberbullying can be defined as the use of information and communication technology by an individual or a group of users to harass other users (Salmivalli, 2010). Cyberbullying has also been extensively recognized as a serious national health problem (Xu, Jun, Zhu, & Bellmore, 2012), in which victims demonstrate a significantly high risk of suicidal ideation (Sampasa-Kanyinga, Roumeliotis, & Xu, 2014). Cyberbullying is a substantially persistent version of traditional forms of bullying with negative effects on the victim. A cyberbully can harass his/her victims before an entire online community. Online social media, such as social networking sites (e.g., Facebook and Twitter) have become integral components of a user's life. Therefore, these websites have become the most common platforms for cyberbullying victimization (Whittaker & Kowalski, 2015), and their popularity and proliferation have increased the incidents of cyberbullying (Mark & Ratliffe, 2011). Such increase is commonly attributed to the fact that traditional bullying is more difficult to practice than cyberbullying, in which the perpetrators bully their victims without direct confrontation by using a laptop or a cellphone connected to the Internet (Kowalski,

* Corresponding author.
E-mail address: Mohammedali@siswa.um.edu.my (M.A. Al-garadi).

http://dx.doi.org/10.1016/j.chb.2016.05.051
0747-5632/

Limber, Limber, & Agatston, 2012). The characteristics of online social networks have also expanded the reach of cyberbullies to previously unreachable locations and countries.

Twitter is a common online social network service that enables users to send and read 140-character messages.[1] The Twitter network currently includes over 500 million users, of which 288 million actively communicate through this network and generate approximately 500 million tweets each day. Approximately 80% of these active Twitter users tweet using their mobile phones. Although this social networking site has become an important, near real-time communication channel (Kavanaugh et al., 2012), a study determined that Twitter is turning into a "cyberbullying playground" (Xu et al., 2012).

In the current research, we aim to utilize useful information in tweets to improve the cyberbullying detection performance. In particular, we use many useful features in Twitter, such as network, activity, user, and tweet content, to train our detection model and improve its performance.

Applying machine learning may provide successful or unsuccessful cyberbullying predication results because building a successful machine learning model depends on many factors (Domingos, 2012). The most important of these factors are the features used and the presence of independent features in the model that correlate well with the class. Selecting the best features with high discriminative power between cyberbullying and non-cyberbullying tweets is a complex task (Domingos, 2012) that requires considerable effort in building the machine learning model (Domingos, 2012). Accordingly, we aim to develop a cyberbullying detection method by identifying discriminative features that can be used in machine learning schemes to distinguish cyberbullying tweets from non-cyberbullying ones. This work provides the following contributions.

- We propose a set of unique features that includes network information, activity information, user information, and tweet content, which are selected based on observations from previous cyberbullying survey research. These observations have been converted to potential features, which are then tested to enhance the discriminative power of the classifiers. As a key novel contribution to the literature, we identify the most significant features and use them as inputs to different machine learning classification algorithms to detect cyberbullying with high accuracy.
- We test different feature combinations and iteratively select different numbers of features to determine a combination that has a significant discriminative power and can obtain improved results. We select three features selection algorithms, namely, χ^2 test, information gain, and Pearson correlation, to determine the most significant proposed features. The synthetic minority oversampling technique (SMOTE) approach and the weights adjusting approach (cost-sensitive) are used to balance the classes in the data set. Thereafter, we compare the performance of four classifiers, namely, naïve Bayes (NB), support vector machine (SVM), random forest, and k-nearest neighbor (KNN), under four different settings to select the best setting for the proposed features.
- Our detection model obtained an area under the receiver operating characteristic (ROC) curve (AUC) of 0.943 and an f-measure of 0.936 using random forest with SMOTE. We compare the result from the proposed features with that from two baseline features, and the comparison outcomes emphasize the significance of the proposed features.

[1] Twitter official website: https://about.twitter.com/company.

2. Related work

A previous study proposed an approach for offensive language detection that was equipped with a lexical syntactic feature and demonstrated a higher precision than the traditional learning-based approach (Chen, Zhou, Zhu, & Xu, 2012). A YouTube data-based study (Dadvar, Trieschnigg, Ordelman, & de Jong, 2013) applied SVM to detect cyberbullying, and determined that incorporating user-based content improved the detection accuracy of SVM. Using data sets from MySpace, *Dadvar* et al. developed a gender-based cyberbullying detection approach that used the gender feature in enhancing the discrimination capacity of a classifier (Dadvar, de Jong, Ordelman, & Trieschnigg, 2012). *Dadvar* et al. and *Ordelman* et al. included age and gender as features in their approach; however, these features were limited to the information provided by users in their online profiles (Dadvar et al., 2012; Dadvar, Trieschnigg, Ordelman, et al., 2013). Moreover, most studies determined that only a few users provided complete information about themselves in their online profiles. Alternatively, the tweet contents of these users were analyzed to determine their age and gender (D. Nguyen, Gravel, Trieschnigg, & Meder, 2013; Peersman, Daelemans, & Van Vaerenbergh, 2011). Several studies on cyberbullying detection utilized profane words as a feature (Kontostathis, Reynolds, Garron, & Edwards, 2013; Reynolds, Kontostathis, & Edwards, 2011), thereby significantly improving the model performance. A recent study (Squicciarini, Rajtmajer, Liu, & Griffin, 2015) proposed a model for detecting cyberbullies in MySpace and recognizing the pairwise interactions between users through which the influence of bullies could spread. *Nalini and Sheela* proposed an approach for detecting cyberbullying messages in Twitter by applying a feature selection weighting scheme (Nalini & Sheela, 2015). *Chavan and Shylaja* included pronouns, skip-gram, TF–IDF, and N-grams as additional features in improving the overall classification accuracy of their model (Chavan & Shylaja, 2015).

However, these features are considered inadequate and are not extensive or discriminative enough to analyze the dynamics of online social network data. Furthermore, the adoption of online social network sites has introduced a new set of acronyms, a few of which are related to cyberbullying. These coded messages may be used as the first clue in detecting cyberbullying engagement. Therefore, including these new acronyms and words can improve the performance of a cyberbullying classifier. Previous studies have not considered several important factors in detecting cyberbullying, such as human behavior and personality (Sanchez & Kumar, 2011). These factors can also be used as features that may increase the discriminative power of the classifier. For example, with respect to the personality of users, survey studies have determined a strong relationship between the personality of a user and cyberbullying engagement (Connolly & O'Moore, 2003; Corcoran, Connolly, & O'Moore, 2012). Most cyberbullies are characterized as neurotic, which is evident in their writing style (Connolly & O'Moore, 2003; Corcoran et al., 2012). Including these characteristics as features can facilitate in gathering clues for cyberbullying detection. Twitter-based cyberbullying detection studies (Bellmore, Calvin, Xu, & Zhu, 2015; Sanchez & Kumar, 2011; Xu et al., 2012) collected their data sets using specific keywords. However, by merely tracking those tweets that contained specific keywords, these studies introduced a potential sampling bias (Cheng & Wicks, 2014; Morstatter, Pfeffer, Liu, & Carley, 2013), limited their detection coverage to such tweets, and disregarded those many other tweets relevant to cyberbullying. These data collection approaches narrow the detection range of cyberbullying. Moreover, the selection of keywords for tracking tweets is subject to the author's perception on cyberbullying. Therefore, the classifiers must be