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/© 2016 Elsevier Ltd. All rights reserved.

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.

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

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

extended from “enriched data” to the complete range of tweets (Xu et al., 2012).

Recent studies have achieved immense success in applying the machine learning technique to detect spam messages in social media (Chen, Chandramouli, & Subbalakshmi, 2014; McCord & Chuah, 2011; Wang, 2010a; Zheng, Zeng, Chen, Yu, & Rong, 2015). However, spam is different from cyberbullying, that is, spam is sent to many people with nearly the same contents, whereas cyberbullying is directed to specific users and has a varying context (Lieberman, Dinakar, & Jones, 2011). Research on cyberbullying detection in social media, such as Twitter, is still in its infancy. We require a cyberbullying tweets detection method with extensive detection coverage, and a substantially accurate and efficient cyberbullying detection method must be proposed to avoid inconvenience for normal tweets. By contrast, tweets contain plenty of information that needs to be utilized to achieve a considerably accurate and efficient cyberbullying detection.

3. Materials and methods

3.1. Data collection

The data were collected from Twitter between January 2015 and February 2015. We focused on geo-tagged tweets within a bounding box in terms of the latitude and longitude of the state of California. These tweets were fetched using the sampled API service of Twitter. Our data set contains 2.5 million geo-tagged tweets. In this study, we extract only the publicly available content through the Twitter API and according to Twitter’s privacy policies to avoid any privacy breach.

The Twitter application program interface (API) enables the researcher to extract public tweets. Each extracted tweet from the Twitter API contains extensive information (Kwak, Lee, Park, & Moon, 2010), such as user ID, username, user biography, user screen name, user URL, user account creation data, tweet text (i.e., the main tweet text containing information about emotions, thoughts, behaviors, and other personally salient information) (Eichstaedt et al., 2015), tweet creation time, tweet’s unique ID, language of the tweet, number of tweets of a user, number of favorites, number of followers, number of mentions, number of following, number of retweets, bounding box of the location (geolocation), and the application that sent the tweet. This information has been used to develop a set of features for efficiently classifying the data from Twitter and used in different applications as well (Bollen, Mao, & Zeng, 2011; Eichstaedt et al., 2015; Java, Song, Finin, & Tseng, 2007; Preotiuc-Pietro et al., 2015; Sakaki, Okazaki, & Matsuo, 2010). We utilize a tweet’s information to develop a proposed set of features and determine the significant features to train our machine-learning model in discriminating between cyberbullying and non-cyberbullying tweets.

The public streaming API of Twitter only provides access to a small sample of relevant data in a few instances, thereby introducing a potential sampling bias (Cheng & Wicks, 2014; González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014; Liu, Kliman-Silver, & Mislove, 2014). Morstatter et al. (2013) analyzed whether the data extracted from this API could sufficiently represent the activities in the Twitter network in general, and determined that when geo-code filtering was used, API would return a nearly complete set of geo-tagged tweets despite the geo-coded sampling process. By contrast, the API would return a data set with certain bias if keyword (i.e., words, phrases, or hashtags) or user ID sampling was adopted. Alternatively, geo-tagged filtering can be applied for collecting data to achieve a favorable representation of the activities in a network. Moreover, researchers who adopt geo-tagged filtering are confident that they are working with an almost complete sample of Twitter data (Morstatter et al., 2013). Therefore, we adopted geo-code filtering to minimize bias in our data set.

The collected data were preprocessed after collection. The tweets were converted to lowercase, www.* or https?//* was converted to URL, @username was converted to AT_USER, additional white spaces were removed, two or more repetitions of characters were replaced by the same character, and misspelled words were corrected using a spelling corrector.

3.2. Feature engineering

We identified four categories of features, namely, network, activity, user, and content, to detect cyberbullying behavior. These features are mainly derived from tweet contents (tweet text) and tweet information, such as network and activity. These features, as explored in the succeeding section, were used in conjunction with a supervised machine learning framework to create a model for cyberbullying detection. We used NB, SVM, random forest, and KNN for machine learning.

All proposed features are investigated (see Results section) using the features analysis technique to determine the best combination of features with the highest discriminative power.

The proposed features are discussed in detail in the following subsections.

3.2.1. Network features

Network features include the number of friends following a user (followers), number of users being followed by a user (following), following–followers ratio, and account verification status. A survey research observed a strong correlation between cyberbullying behavior and sociability of users in online environments (Navarro & Jasinski, 2012). Therefore, these features measure the sociability of a user in Twitter (Lee, Mahmud, Chen, Zhou, & Nichols, 2014).

3.2.2. Activity features

Activity features measure the online communication activity of a user (Pennacchiotti & Popescu, 2011). The number of posted tweets, favorited tweets, and URLs, hashtags, and mentioned users (e.g., AT_USERname) in a tweet was extracted to measure the activeness of users in Twitter. A survey research determined that those users who were considerably active in online environments were likely to engage in cyberbullying behavior (Balakrishnan, 2015).

3.2.3. User features

3.2.3.1. Personality feature. Online social networks have become a place where users can present and introduce themselves to the virtual world. Many researchers have utilized online social network data to predict the personality of users within such networks (Adali & Golbeck, 2012; Golbeck, Robles, & Turner, 2011; Quercia, Kosinski, Stillwell, & Crowcroft, 2011). In particular, these researchers predict personality by analyzing online communication data from social media, and personality prediction in social media can facilitate the understanding of human behaviors (Mahmud, Zhou, Megiddo, Nichols, & Drews, 2013).

Cyberbullying is associated with the aggressive behaviors of a user. Survey studies show that hostility significantly predicts cyberbullying (Arıcak, 2009), and that both bullying and cyberbullying are strongly related to neuroticism (Connolly & O’Moore, 2003; Corcoran et al., 2012). Neuroticism is characterized as anxiety, anger, and moodiness. Neurotic people are angrier, moodier, and more tense than normal people, thereby indicating that a neurotic user is likely to engage in cyberbullying. Therefore, predicting neurotic personality and neurotic-related text can provide useful discriminative information.

The words in one's writings, such as blogs and essays, are also related to his/her personality (Fast & Funder, 2008; Gill, Nowson, & Oberlander, 2009; Mairesse & Walker; Tausczik & Pennebaker, 2010). Previous studies (Golbeck, Robles, Edmondson, & Turner, 2011; Golbeck, Robles, & Turner, 2011; Mahmud et al., 2013) encourage the use of social media text for personality prediction even if the text contains only a few words. Previous studies on personality prediction in social media reveal a positive correlation between neuroticism and usage of anxiety- and anger-related words (Adali & Golbeck, 2012; Golbeck, Robles, & Turner, 2011; Quercia et al., 2011; Schwartz et al., 2013). Therefore, predicting neuroticism can provide a beneficial discriminative feature in detecting cyberbullying behavior. To predict neuroticism, we used the one hundred most common words used in social media, which are positively correlated with neuroticism and the one hundred most common words used in social media, which are negatively correlated with neuroticism, as introduced in study (Schwartz et al., 2013).

3.2.3.2. Gender. Many researchers have investigated the relationship between gender and engagement in cyberbullying behaviors. A few studies (Calvete, Orue, Estévez, Villardón, & Padilla, 2010; Vandebosch & Van Cleemput, 2009) show that males are more likely to engage in cyberbullying behaviors than females, but other studies show the opposite (Dilmac, 2009; Sourander et al., 2010). Moreover, another study determine no significant difference between males and females in terms of tendency to engage in cyberbullying behaviors (Kowalski, Giumetti, Schroeder, & Reese, 2012). Even though the survey research do not clearly confirm the relationship between gender and cyberbullying behavior as previously explained, the study (Van Royen, Poels, Daelemans, & Vandebosch, 2015) that conducted a survey involving numerous experts in the field of cyberbullying suggested to include gender to build effective machine learning models (Van Royen et al., 2015). Similarly, the study showed that males used cyberbullying to a considerable extent than females do (Calvete et al., 2010). Moreover, a study determined that accurate cyberbullying detection in online social networks was improved by using gender-related information as a feature (Dadvar et al., 2012). Therefore, these observations prompted us to include gender-related features in building our machine-learning model.

Unfortunately, most users do not mention their gender in their profiles, and not all gender information that is provided in user profiles are correct (Peersman et al., 2011). However, recent studies have applied natural language processing to predict the gender of users based on their writing styles. Previous studies show that males and females use specific words to distinguish themselves from the opposite gender. For example, females use the word "shopping" more frequently than males (Schwartz et al., 2013). Accordingly, we used several features to predict the gender of a user. First, we predicted gender based on the tweet text (Schwartz et al., 2013), that is, we used the 100 most common words that are used/not used by males/females (Schwartz et al., 2013). Second, we predicted gender based on the first name of the user (Liu & Ruths, 2013). In particular, we used a large gender-labeled data set and predicted whether a user was male or female based on the first name reported in his/her tweet.

3.2.3.3. Age. Previous studies have discussed the effects of age on engagement in cyberbullying behavior. These studies contend that cyberbullying behavior decreases along with increasing age, and that the rate of cyberbullying behavior is highest among teenage users (Slonje & Smith, 2008; Williams & Guerra, 2007). However, the older age group must still be considered. Most online social network users do not provide information on their age or date of birth, and not all users provide accurate age information (Peersman et al., 2011). Similar to gender, the lack of information about age imposes a challenge to this study. However, user age can be predicted by analyzing the language used by users from different age levels. We used the same age levels reported in (Schwartz et al., 2013) (i.e., age 1: 13 years–18 years; age 2: 19 years–22 years; age 3: 23 years–29 years; and age 4: 30 years and above), and the 100 most common words used in social media that were positively/negatively correlated with each age level (Schwartz et al., 2013). We used nearly 800 age-related words in social media to distinguish users from different age levels. For example, the word "school" is significantly related to age 1, whereas "job" is substantially related to age 3.

Note that as mentioned above, most OSN users do not clearly state their age and gender in their online profiles (Peersman et al., 2011); thus, many studies (Hosseini & Tammimy, 2016; Peersman et al., 2011; Rangel & Rosso, 2013; Santosh, Bansal, Shekhar, & Varma, 2013; Talebi & Kose, 2013) have focused on predicting age and gender in OSNs by using text analysis and natural language processing on a user's posts. The aforementioned study (Schwartz et al., 2013) has developed an open vocabulary by analyzing 700 million words, phrases, and topic instances collected from social media and determined the words related to age, gender, and personality. The study (Schwartz et al., 2013) provided the comprehensive exploration of language that distinguishes people, thereby determining connections that are not obtained with the traditional closed vocabulary word category analysis, such as Linguistic Inquiry and Word Count (LIWC). The proposed open vocabulary related to gender and age has been used in various studies. For example, a study used these vocabularies to predict county-level heart disease mortality on Twitter (Eichstaedt et al., 2015). Another study used these vocabularies to propose a feature related to age and gender to detect mental illnesses, such as depression and post-traumatic stress disorder, on Twitter (Preotiuc-Pietro et al., 2015). These vocabularies were also used for predicating age and gender features from a user's posts on social media websites to build a machine learning classifier for different applications (Burger, Henderson, Kim, & Zarrella, 2011; Li, Sun, & Liu, 2014; Miller, Dickinson, & Hu, 2012; D.-P. Nguyen, Gravel, Trieschnigg, & Meder, 2013; Rangel & Rosso, 2013; Rao, Yarowsky, Shreevats, & Gupta, 2010). In our study, we used the open vocabulary approach related to age and gender proposed in the aforementioned study (Schwartz et al., 2013). This open vocabulary was used in our study as a feature to predicate gender and age of tweet authors from the information provided by their tweets.

3.2.4. Content features

3.2.4.1. Vulgarities feature. Using vulgar words in online communication can be used to detect cyberbullying because they may signal hostility and offensive behavior. Moreover, tweets containing a vulgar or profane word may be considered a cyberbullying tweet (Xiang, Fan, Wang, Hong, & Rose, 2012).

Vulgarity is a useful discriminative feature for detecting offensive and cursing behavior in Twitter (Wang, Chen, Thirunarayan, & Sheth, 2014; Xiang et al., 2012) and cyberbullying in YouTube comments (Dadvar, Trieschnigg, Ordelman, et al., 2013). These features are extracted from the contents of a user's post. We measured the number of profane words in a post using a dictionary of profanity. The words in this dictionary were compiled from the sources cited in previous studies (Reynolds et al., 2011; Wang et al., 2014).

3.2.4.2. Specific social network cyberbullying features. Technology-mediated communication has immensely contributed to the increasing number of novel acronyms and abbreviations. By introducing new terms, such as “unfriend” and “selfie,” social media evidently have an effect on language. The acronyms and abbreviations that are generated from online social media communication assist users communicate easily and rapidly with one another. These terms can also be easily entered in mobile phones with tiny keypads. On Twitter, acronyms assist users make the most out of 140 characters. Acronyms and abbreviations have also become popular among mobile phone users who use these terms to reduce their effort and time for typing. By introducing new words, adding new meanings to old words and changing the manner we communicate, social media have made their presence.

Similarly, cyberbullies change the method they use words and acronyms to engage in cyberbullying. Online social networks facilitate in creating cyberbullying-related acronyms that have never been used in traditional bullying or beyond social media. The words, acronyms, and abbreviations commonly used in cyberbullying were collected from (Dailymail, 2014).

3.2.4.3. First and second person. First and second person pronouns in tweets are also used as features that provide useful information about to whom the text is directed. A text containing cyberbullying-related features and a second person pronoun is most likely to be meant for harassing others (Dadvar, Trieschnigg, & Jong, 2013).

3.3. Machine learning algorithms

The features extracted from the tweets were used to construct a model for detecting cyberbullying behavior. We tested several machine learning techniques to select the best classifier. We tested NB, SVM (LibSVM), decision trees (DT) (random forest), and KNN (Hall et al., 2009) using the WEKA tool (Hall et al., 2009).

These techniques are discussed as follows.

3.3.1. Naïve Bayes (NB)

NB is considered one of the most efficient and effective inductive learning algorithms in the field of machine learning (Zhang, 2004a), and has been used as an effective classifier in several studies on social media (Bora, Zaytsev, Chang, & Maheswaran, 2013; Freeman, 2013; Wang, 2010b). NB is a classification algorithm that is based on the application of the Bayes theorem with strong (naive) independence assumptions. Given a class variable $\boldsymbol{y}$ and a dependent feature vector $\boldsymbol{x}_1$ through $\boldsymbol{x_n}$, Bayes' theorem states the following relationship:

$$p(y|x_1,\ x_2,\cdots,\ x_n\ ) = \frac{p(y)p(x_1,\ x_2,\cdots,\ x_n\ |y)}{p(x_1,\ x_2,\cdots,\ x_n)}. \tag{1}$$

Using the naive independence assumption, we obtain the following equation:

$$p(x_i|y,x_1,\cdots x_{i-1},x_i+1,\cdots,\ x_n) = p(x_i|y), \tag{2}$$

for all i, this relationship is simplified as follows:

$$p(y|(x_1,\ x_2,\ \cdots,\ x_n\ ) = \frac{p(\mathrm{y})\prod_{i=1}^{\mathrm{n}} p(x_i|y)}{p(x_1,\ x_2,\ \cdots,\ x_n)}. \tag{3}$$

Given that $p(y|$ is a constant because of the input, we use the following classification rule:

$$p(y|(x_1,\ x_2,\ \cdots,\ x_n)\propto p(\mathrm{y})\prod_{i=1}^{\mathrm{n}} p(x_i|y). \tag{4}$$

This technique is discussed in further detail in (Zhang, 2004b).

3.3.2. Support vector machine (SVM)

SVM is a popular supervised method that is based on statistical learning theories (Vapnik, 2000) and is commonly used for detecting anomaly and network intrusion. For example, we consider an SVM binary classification. Given a training data $(x_1,\ x_2,\ \cdots,\ x_n)$, which are expressed as vectors in a certain space $X \subseteq$Rd. These data are labeled $(\mathrm{y}_1 \ldots \mathrm{y}_\mathrm{m})$, where $y_i \in (-1,\ 1)$. In their simplest form, SVMs are hyperplanes that separate the training data by a maximal margin. Class (-1) is on one side of the hyperplane, whereas class (1) is on the other side. In our case, cyberbullying is on one side of the hyperplane, whereas non-cyberbullying is on the other side.

3.3.3. Random forest

Random forest is an ensemble learning technique that builds multitudes of decision tress. A decision tree is a graphic method in which each branch node represents a choice between alternatives. A graphic approach is employed in decision trees to compare competing alternatives.

3.3.4. KNN

KNN is a non-parametric method that attempts to determine the K nearest neighbors of x_0 and uses a majority vote to determine the class label of x_0. Without prior knowledge, the KNN classifier often applies Euclidean distances as the distance metric.

3.4. Manual dataset annotation

Cyberbullying behaviors are defined as aggressive behaviors exhibited through electronic or digital media and intended to inflict harm or discomfort to a victim (Bauman, Toomey, & Walker, 2013) (Kowalski, Giumetti, Schroeder, & Lattanner, 2014). According to the critical review and meta-analysis of cyberbullying (Kowalski et al., 2014), most researchers agree that cyberbullying involves the use of electronic communication technologies to bully others. Cyberbullying can take many forms, including posting hostile comments, frightening or harassing the victim, producing hateful or insulting posts, or abusing the victim (Q. Li, 2007; Tokunaga, 2010).

We randomly selected 10,606 tweets from our collected data, and each tweet was manually classified as cyberbullying or non-cyberbullying. Generally the tweets are labeled via human coding (Dadvar & De Jong, 2012; Diakopoulos & Shamma, 2010; Ghahramani, 2015; Gokulakrishnan, Priyanthan, Ragavan, Prasath, & Perera, 2012; Myslín, Zhu, Chapman, & Conway, 2013; Paul, Dredze, & Broniatowski, 2014; Prieto, Matos, Alvarez, Cacheda, & Oliveira, 2014; Räbiger & Spiliopoulou, 2015). In this study, the tweets were labeled with the assistance of three experts based on the abovementioned cyberbullying definition. Moreover, these experts were oriented about the abbreviations, slang words, and acronyms commonly used in social network and online communications to assist them further understand the tweet contents. The tweets were classified as follows.

- Cyberbullying: the tweet content indicates the presence of cyberbullying behaviors.
- Non-cyberbullying: the tweet content does not indicate the presence of cyberbullying behavior.

The tweets are considered cyberbullying if at least two of the assigned experts consider such as cyberbullying tweets. If these experts do not agree on the classification of a tweet, this tweet will be deleted from the data set. Our final data set contains 10,007 tweets that are classified as non-cyberbullying (9408) and cyberbullying (599).

We used experts to code the tweets instead of using human coding through Mechanical Turk website (M.Turk) to improve the quality of the labeling processes. Using experts would also avoid the online spam trucker, who simply classifies the tweets without actually reading them (Ipeirotis, 2010). However, using experts was more time consuming than M. Turk.

3.5. Feature analysis and selection

Feature analysis is performed to determine the most significant features. This method can lessen the classification time and remove the redundant features that do not provide favorable discriminative patterns. Not all features are expected to be significant and contributive to the model results; thus, three feature selection algorithms were computed, namely, χ^2 test (chi-square test), information gain, and Pearson correlation, to determine the discriminative power of each feature (Yang & Pedersen). We combined different features to determine a combination with a significant discriminative power and can provide improved results. The top ten significant features selected by abovementioned feature selection methods are presented in the Table 1.

3.6. Handling imbalanced class distribution

Our data set shows an imbalanced class distribution, that is, only 599 tweets are classified as cyberbullying, whereas 10,007 tweets are classified as non-cyberbullying. Such imbalanced class distribution can prevent the model from accurately classifying the instances. Learning algorithms that lack class imbalance tend to be overwhelmed by the major class and ignore the minor one. In real-world applications, data sets often contain imbalanced data in which the normal class forms the majority and the abnormal class forms the minority, such as fraud detection, instruction detection, and medical diagnosis. Several approaches for overcoming these problems have been proposed, such as a combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), as well as weight adjusting approaches (Liu & Zhou, 2006). Accordingly, we employed both approaches in the current study. The next section presents the comparative results before and after using each approach.

3.7. Experiments

We ran an extensive set of experiments to measure the performance of the four classifiers (e.g., NB, LIBSVM, random forest, and KNN).

3.7.1. Settings

All four classifiers were tested in four different settings, namely, basic classifiers, classifiers with feature selection techniques, classifiers with SMOTE alone and with feature selection techniques, and classifiers with cost-sensitive alone and with feature selection techniques. All experiments were based on a 10-fold cross validation (Kohavi, 1995; Refaeilzadeh, Tang, & Liu, 2009). First, we tested these classifiers using all of the proposed features. Second, we tested these classifiers with feature selections (i.e., χ^2, information gain, and Pearson correlation). Third, we tested these classifiers with SMOTE alone and with feature selection techniques. Fourth, we tested these classifiers with cost-sensitive alone and with feature selection techniques.

4. Results

Our data set contains an imbalanced class distribution; therefore, the selection of an evaluation metric is considerably important. We selected AUC as our main performance measure because of its high robustness for evaluating classifiers.

AUC has a significant statistical property. Moreover, AUC of a classifier is equal to the probability that the classifier will rank a randomly selected positive instance higher than a randomly selected negative instance (Fawcett, 2006). AUC (the area under the ROC Curve) is commonly used in medical decision-making and have been used intensively in recent years in machine learning and data-mining studies, where the imbalanced data class exists (Fawcett, 2006). The key advantage of AUC is that it is more robust than accuracy, precision, recall, and f-measure in class imbalanced situations. Given a 95% imbalance (e.g., in favor of the positive class), the accuracy of the default classifier that consistently issues “positive” will be 95%, whereas a considerably interesting classifier that actually deals with the issue is likely to obtain a worse score. The ROC curve denotes the rate of true positives versus false positives at different threshold settings. The area under the curve provides a signal of the discriminatory rate of the classifier at various operating points (Fawcett, 2004, 2006; Prieto et al., 2014; Provost & Fawcett, 1997). Therefore, AUC gives a robust classifier performance measurement under imbalance class distribution data. A high AUC indicates an improved classification for both class instances regardless of class imbalance (Fawcett, 2006). We also report precision, recall, and f-measure as reference measures.

Table 1
Top ten significant features selected by chi-square test, information gain, and Pearson correlation.

χ^2 test (chi-square test)	Information gain	Pearson correlation
Vulgarities feature (number of vulgar words in the post).	Vulgarities feature (number of vulgar words in the post).	Vulgarities feature (number of vulgar words in the post).
100 most commonly used words in social media that are positively correlated with neuroticism	100 most commonly used words in social media that are positively correlated with neuroticism	100 most commonly used words in social media that are positively correlated with neuroticism
100 most commonly used words in social media that are used by males	100 most commonly used words in social media that are used by males	100 most commonly used words in social media that are used by males
Average number of followers to following	100 most commonly used words in social media that negatively correlate with age (30 years and above)	100 most commonly used words in social media that positively correlate with age (19–22 years)
100 most commonly used words in social media that positively correlate with age (19–22 years)	100 most commonly used words in social media that positively correlate with age (19–22 years)	100 most commonly used words in social media that negatively correlate with age (30 years and above)
100 most commonly used words in social media that negatively correlate with age (30 years and above)	Number of tweets	Number of tweets
Number of friends following a user	Average number of followers to following	Number of mentions
Number of tweets	Second person pronouns	Second person pronouns
Second person pronouns	Number of friends following a user	Average number of followers to following
Number of mentions	Number of mentions	Slang feature (number of slang words in the post)

4.1. Results obtained by using basic classifiers

All four classifiers were run using all proposed features based on a 10-fold cross validation. Table 2 shows the results for each classifier. The AUC results varied between 0.5 and 0.69. NB showed the best overall performance under a basic setting with an f-measure varying between 0.903 and 0.920.

4.2. Result obtained by classifiers with feature selection

We ran all four classifiers with feature selection to determine the most significant feature that may improve the performance of the classifier and reduce classification time. Three feature selection algorithms, namely, χ^2 test, information gain, and Pearson correlation, were tested in the experiment. Different feature combinations were tested and different numbers of features were iteratively selected to determine a combination with a significant discriminative power and could provide an improved result. Tables 3–5 compare the four classifiers with each feature selection method.

Compared with Table 1,Table 2 shows that the results obtained by using χ^2 test to select the significant features has slightly improved AUC for NB (0.704) and random forest (0.629), but slightly decreased the AUC for KNN (0.568). AUC for SVM (0.500) remained the same.

Table 4 shows the results obtained by using the information gain to select the significant features. Similar to the χ^2 test, AUC for NB (0.705) and random forest (0.637) was slightly improved compared to the results in Table 2, which used all features to train the model. Meanwhile, AUC for KNN (0.570) was slightly reduced, whereas that for SVM (0.500) remained the same.

Table 5 shows the results obtained by using the Pearson correlation to select the significant features. Compared with the results in Table 2, AUC for NB (0.701) and random forest (0.646) was slightly improved, whereas that for SVM (0.500) and KNN (0.588) remained the same.

In summary, using the three feature selection techniques has only slightly improved the AUC results compared with using all features to train the model Table 2.

4.3. Result obtained by classifier with imbalanced data distribution

We over-sampled the minority (abnormal) class and under-sampled the majority (normal) class (SMOTE), as well as applied weights adjusting approaches (cost-sensitive), to handle the imbalanced data distribution. We ran four classifiers with these two approaches with and without feature selection to obtain the best result.

Table 6 shows the results obtained by using the four classifiers with SMOTE. In summary, using SMOTE (300%) significantly improved AUC for all classifiers, except for NB, which only showed a small improvement.

Table 7 shows the results obtained using the four classifiers with the weights adjusting approach (cost-sensitive). In summary, using classifiers with cost-sensitive did not significantly improve AUC of the classifiers.

Table 2
Results obtained by using basic classifiers.

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.897	0.903	0.690
LibSVM	0.890	0.944	0.916	0.500
Random forest	0.908	0.942	0.917	0.626
KNN	0.910	0.937	0.920	0.588

Table 3
Result obtained by using the χ^2 test (chi-square test).

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.901	0.905	0.704
LibSVM	0.890	0.943	0.916	0.500
Random forest	0.903	0.940	0.917	0.629
KNN	0.907	0.935	0.918	0.568

Table 4
Result obtained by using the information gain.

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.901	0.904	0.705
LibSVM	0.944	0.890	0.943	0.500
Random forest	0.904	0.940	0.917	0.637
KNN	0.904	0.933	0.916	0.570

Table 5
Results obtained by using the Pearson correlation.

Classifier	Precision	Recall	F-measure	AUC
NB	0.909	0.898	0.904	0.701
LibSVM	0.890	0.944	0.916	0.500
Random forest	0.901	0.941	0.916	0.646
KNN	0.910	0.937	0.920	0.588

4.4. Results summary

Table 6 shows that the best overall classifiers performance was achieved using SMOTE. In particular, random forest using SMOTE alone showed the best AUC (0.943) and f-measure (0.936).

The following figures compare the ROC results of all classifiers under the basic and best performance settings (classifiers using SMOTE). (See Figs. 1 and 2).

The confusion table for the best performance classifier Random forest using SMOTE is presented in Table 8.where

- True positive (TP) is the percentage of instances that are non-cyberbullying and correctly classified as non-cyberbullying.
- False negative (FN) is the percentage of instances that are non-cyberbullying and incorrectly classified as cyberbullying.
- True negative (TN) is the percentage of instances that are cyberbullying and correctly classified as cyberbullying.
- False positive (FP) is the percentage of instances that are cyberbullying and incorrectly classified as non-cyberbullying

Overall, the machine learning working in an online communication environment should be balanced between providing effective detecting models to detect cyberbullying content or any negative behavior that does not ethically harm other innocent contents or users. The preceding table shows that the proposed machine learning model classified 99.4% of non-cyberbullying as non-cyberbullying; therefore, this result meets (Vayena, Salathé, Madoff, Brownstein, & Bourne, 2015) the ethical challenges because the content, which was falsely detected as cyberbullying but was actually not, is at a low rate of 0.6% of non-cyberbullying tweets classified as cyberbullying. By contrast, the machine-learning model classified 71.4% of cyberbullying as cyberbullying, thereby indicating that it is an effective in detecting cyberbullying. Therefore, the performance results (AUC = 0.943) and confusion table of the Random forest using SMOTE emphasize that the proposed model based on the proposed features provides a feasible solution in detecting cyberbullying in online communication environments. (See Table 6).

Table 6
Result obtained using SMOTE.

Cases	Classifier	Precision	Recall	F-measure	AUC
SMOTE only	NB	0.763	0.774	0.768	0.692
	LibSVM	0.820	0.831	0.786	0.583
	Random forest	**0.941**	**0.939**	**0.936**	**0.943**
	KNN	0.870	0.873	0.871	0.866
SMOTE and χ^2 test	NB	0.760	0.769	0.764	0.703
	LibSVM	0.820	0.833	0.792	0.593
	Random forest	0.934	0.934	0.930	0.924
	KNN	0.863	0.871	0.865	0.846
SMOTE and information gain	NB	0.764	0.774	0.769	0.717
	LibSVM	0.823	0.835	0.796	0.599
	Random forest	0.897	0.897	0.897	0.929
	KNN	0.861	0.869	0.863	0.843
SMOTE and Pearson correlation	NB	0.762	0.781	0.770	0.708
	LibSVM	0.829	0.836	0.796	0.598
	Random forest	0.939	0.938	0.934	0.932
	KNN	0.864	0.871	0.866	0.850

Table 7
Result obtained using cost-sensitive.

Cases	Classifier	Precision	Recall	F-measure	AUC
Cost-sensitive only	NB	0.909	0.897	0.903	0.690
	LIBSVM	0.919	0.944	0.916	0.502
	Random forest	0.908	0.942	0.917	0.626
	KNN	0.910	0.937	0.920	0.588
Cost-sensitive and χ^2 test	NB	0.909	0.901	0.905	0.704
	LIBSVM	0.919	0.944	0.916	0.502
	Random forest	0.903	0.940	0.917	0.629
	KNN	0.907	0.935	0.918	0.568
Cost-sensitive and information gain	NB	0.909	0.901	0.904	0.705
	LIBSVM	0.913	0.943	0.916	0.502
	Random forest	0.904	0.940	0.917	0.637
	KNN	0.904	0.933	0.916	0.570
Cost-sensitive and Pearson correlation	NB	0.909	0.898	0.904	0.701
	LIBSVM	0.947	0.944	0.917	0.502
	Random forest	0.930	0.901	0.941	0.646
	KNN	0.912	0.938	0.921	0.573

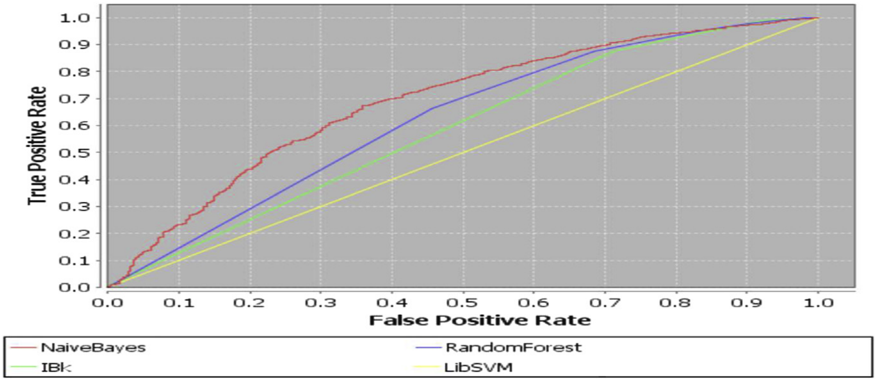

Fig. 1. ROC results for the four classifiers under the basic setting.

5. Significance of the proposed feature sets

Given the restrictions in API and the possible ethical/privacy considerations, no public twitter data set was available to test the significance of our proposed feature sets. To investigate such significance, we created two baseline features from our data set, namely, a bag of words and a combination of possible features proposed in previous studies (Chavan & Shylaja, 2015; Dadvar et al., 2012; Dadvar, Trieschnigg, Ordelman, et al., 2013; Kontostathis et al., 2013; Reynolds et al., 2011). We ran an extensive set of experiments to measure the performance of four classifiers using these two baseline features under different settings. The first baseline feature achieved the best result for NB with information gain (AUC = 0.614), whereas the second baseline feature achieved the best result for random forest using SMOTE alone (AUC = 0.724). We compared the best result obtained from the proposed features with those obtained from two baseline features. Table 9 shows the significance of the proposed features.

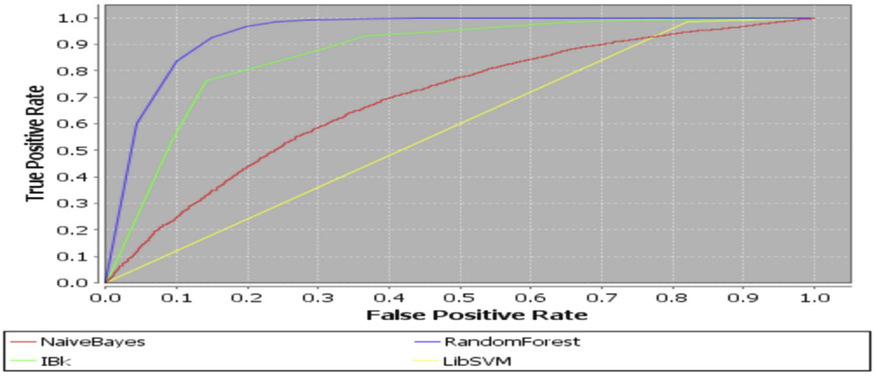

Fig. 2. ROC results for the four classifiers using SMOTE alone.

Table 8
Confusion table.

		Classified	
		Non-cyberbullying	Cyberbullying
Actual	Non-Cyberbullying	99.4% (T.P)	0.6% (F.N)
	Cyberbullying	28.6% (F.P)	71.4% (T.N)

Table 9
Comparison of the AUC results of the proposed and baseline features.

Features used	AUC under the best setting
Proposed features	0.943
Baseline 1	0.614
Baseline 2	0.724

6. Conclusion

Cyberbullying has become a major problem along with the development of online communication and social media. Committed by an individual or a group of users, cyberbullying refers to the use of information and communication technology to harass others. This phenomenon has been extensively acknowledged as a serious national health problem in which victims demonstrate a significantly high risk of suicidal ideation.

We developed a model for detecting cyberbullying in Twitter. The developed model is a feature-based model that uses features from tweets, such as network, activity, user, and tweet content, to develop a machine learning classifier for classifying the tweets as cyberbullying or non-cyberbullying.

We ran an extensive set of experiments to measure the performance of the four selected classifiers, namely, NB, LibSVM, random forest, and KNN. Three features selection algorithms were selected, namely, χ^2 test, information gain, and Pearson correlation, to determine the most significant feature. Feature analysis algorithms were applied using different feature combinations, and different numbers of features were iteratively selected to determine a combination with a significant discriminative power and could provide an improved result. Over-sampling of the minority (abnormal) class and under-sampling of the majority (normal) class (SMOTE) were applied along with weights adjusting approaches (cost-sensitive) to handle the imbalanced class distribution in our manually labeled data set. SMOTE improved the overall performance of the classifiers. Given that our manually labeled dataset contains imbalanced class distribution, AUC was used as our main performance measure because of its high robustness for evaluating classifiers. The best overall classifiers performance was achieved by using classifiers that use SMOTE to handle the imbalanced data distribution. Random forest using SMOTE alone showed the best AUC (0.943) and f-measure (0.936). The comparison between the best results from our proposed features and those from the two baseline features emphasize the significance of our proposed features.

The proposed model can be used by the organization's members, such as parents, guardians, educational institutions, and organizations (e.g., workplace), as well as non-government organizations (NGOs), including crime-prevention foundations, social chamber organizations, psychiatric associations, policy makers, and enforcement bodies.

Individual behavior and mood are an affective state that is important for physical and emotional well-being, creativity, and decision-making (Golder & Macy, 2011; Ruths & Pfeffer, 2014). Using Twitter API, the study in Golder & Macy (2011)) analyzed changes in hourly, daily, and seasonal affect at individual levels. The authors in Golder & Macy (2011) have investigated positive affect (PA), such as enthusiasm, delight, and activeness; and negative affect (NA), such as distress, fear, anger, guilt, and disgust. Therefore, the positive affect varies with seasonal variation (Golder & Macy, 2011). Similarly, an interesting research area is investigating how the seasonal variation of user's mood and psychological condition during the year can affect the language used to exhibit cyberbullying behavior and if this change can occur, how it will affect the accuracy of the machine learning detection. Collecting long-term data will enable machine learning to be trained using a variety of human behavior data with the psychological conditions of different users. Future studies may also use data from other social media to investigate cyberbullying behaviors. A social networking graph that represents the relationships among users can also be used to model this problem.

Acknowledgements

We would like to take this opportunity to thank University of Malaya UMRG (RP028D-14AET) for funding this Research.

References

Adali, S., & Golbeck, J. (2012). Predicting personality with social behavior. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (pp. 302–309). IEEE Computer Society.
Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). PhishAri: automatic real-time phishing detection on twitter. In *eCrime Researchers Summit (eCrime), 2012* (pp. 1–12). IEEE.
Arıcak, O. T. (2009). Psychiatric symptomatology as a predictor of cyberbullying among university students. *Eurasian Journal of Educational Research, 34*(1), 169.
Balakrishnan, V. (2015). Cyberbullying among young adults in Malaysia: the roles of gender, age and internet frequency. *Computers in Human Behavior, 46*, 149–157.

Bauman, S., Toomey, R. B., & Walker, J. L. (2013). Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence, 36*(2), 341–350.

Bellmore, A., Calvin, A. J., Xu, J.-M., & Zhu, X. (2015). The five W's of "bullying" on Twitter: who, what, why, where, and when. *Computers in Human Behavior, 44*, 305–314.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Bora, N., Zaytsev, V., Chang, Y.-H., & Maheswaran, R. (2013). Gang networks, neighborhoods and holidays: spatiotemporal patterns in social media. In *Social Computing (SocialCom), 2013 International Conference on (pp. 93–101)*. IEEE.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on empirical methods in natural language processing* (pp. 1301–1309). Association for Computational Linguistics.

Calvete, E., Orue, I., Estévez, A., Villardón, L., & Padilla, P. (2010). Cyberbullying in adolescents: modalities and aggressors' profile. *Computers in Human Behavior, 26*(5), 1128–1135.

Chavan, V. S., & Shylaja, S. (2015). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *Advances in computing, communications and informatics (ICACCI), 2015 International Conference on (pp. 2354–2358)*. IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*(1), 321–357.

Chen, X., Chandramouli, R., & Subbalakshmi, K. P. (2014). Scam detection in Twitter. In *Data mining for service* (pp. 133–150). Springer.

Cheng, T., & Wicks, T. (2014). Event detection using Twitter: a spatio-temporal approach. *PLoS One, 9*(6), e97807.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, security, risk and trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)* (pp. 71–80). IEEE.

Connolly, I., & O'Moore, M. (2003). Personality and family relations of children who bully. *Personality and Individual Differences, 35*(3), 559–567.

Corcoran, L., Connolly, I., & O'Moore, M. (2012). Cyberbullying in Irish schools: an investigation of personality and self-concept. *The Irish Journal of Psychology, 33*(4), 153–165.

Dadvar, M., & De Jong, F. (2012). Cyberbullying detection: a step toward a safer Internet yard. In *Proceedings of the 21st International Conference companion on world wide web* (pp. 121–126). ACM.

Dadvar, M., de Jong, F., Ordelman, R., & Trieschnigg, R. (2012). *Improved cyberbullying detection using gender information*.

Dadvar, M., Trieschnigg, D., & Jong, F. (2013a). *Expert knowledge for automatic detection of bullies in social networks*.

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013b). Improving cyberbullying detection with user context. In *Advances in information retrieval* (pp. 693–696). Springer.

Dailymail. (2014). *From IHML (I hate my life) to Mos (mum over shoulder): Why this guide to cyber-bullying slang may save your child's life*.

Diakopoulos, N. A., & Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on human factors in computing systems* (pp. 1195–1198). ACM.

Dilmac, B. (2009). Psychological needs as a predictor of cyber bullying: a preliminary report on college students. *Educational Sciences: Theory and Practice, 9*(3), 1307–1325.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., et al. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science, 26*(2), 159–169.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94*(2), 334.

Fawcett, T. (2004). ROC graphs: notes and practical considerations for researchers. *Machine Learning, 31*(1), 1–38.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Freeman, D. M. (2013). Using naive bayes to detect spammy names in social networks. In *Proceedings of the 2013 ACM workshop on artificial intelligence and security* (pp. 3–12). ACM.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452–459.

Gill, A. J., Nowson, S., & Oberlander, J. (2009). *What are they blogging about? Personality, topic and motivation in blogs*. ICWSM.

Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012). Opinion mining and sentiment analysis on a twitter data stream. In *Advances in ICT for emerging regions (ICTer), 2012 International Conference on (pp. 182–188)*. IEEE.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. In , *IEEE Third International Conference on (pp. 149–156)Privacy, security, risk and trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011*. IEEE.

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 253–262). ACM.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333*(6051), 1878–1881.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks, 38*, 16–27.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

Hosseini, M., & Tammimy, Z. (2016). Recognizing users gender in social media using linguistic features. *Computers in Human Behavior, 56*, 192–197.

Ipeirotis, P. (2010). *Mechanical turk: Now with 40.92% spam*. Behind Enemy Lines blog.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis* (pp. 56–65). ACM.

Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., et al. (2012). Social media use by government: from the routine to the critical. *Government Information Quarterly, 29*(4), 480–491.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013). Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 195–204). ACM.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin, 140*(4), 1073.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Reese, H. H. (2012). Cyber bullying among college students: evidence from multiple domains of college life. *Cutting-edge Technologies in Higher Education, 5*, 293–321.

Kowalski, R. M., Limber, S., Limber, S. P., & Agatston, P. W. (2012). *Cyberbullying: Bullying in the digital age*. John Wiley & Sons.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. In *Proceedings of the 19th International Conference on world wide web* (pp. 591–600). ACM.

Lauw, H. W., Shafer, J. C., Agrawal, R., & Ntoulas, A. (2010). Homophily in the digital world: a LiveJournal case study. *Internet Computing, IEEE, 14*(2), 15–23.

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2014). Who will retweet thisfi. In *Proceedings of the 19th International Conference on intelligent user interfaces* (pp. 247–256). ACM.

Li, Q. (2007). New bottle but old wine: a research of cyberbullying in schools. *Computers in Human Behavior, 23*(4), 1777–1791.

Lieberman, H., Dinakar, K., & Jones, B. (2011). Let's gang up on cyberbullying. *Computer, 44*(9), 93–96.

Li, L., Sun, M., & Liu, Z. (2014). Discriminating gender on Chinese microblog: a study of online behaviour, writing style and preferred vocabulary. In *Natural computation (ICNC), 2014 10th International Conference on (pp. 812–817)*. IEEE.

Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). The tweets they are a-changin': evolution of twitter users and behavior. In *International AAAI Conference on weblogs and social media (ICWSM)* (Vol. 13, p. 55).

Liu, W., & Ruths, D. (2013). In *What's in a name? Using first names as features for gender inference in Twitter*.

Liu, X.-Y., & Zhou, Z.-H. (2006). The influence of class imbalance on cost-sensitive learning: an empirical study. In , *ICDM'06. Sixth international Conference on (pp. 970–974)Data mining, 2006*. IEEE.

Mahmud, J., Zhou, M. X., Megiddo, N., Nichols, J., & Drews, C. (2013). Recommending targeted strangers from whom to solicit information on social media. In *Proceedings of the 2013 International Conference on intelligent user interfaces* (pp. 37–48). ACM.

Mairesse, F., & Walker, M.. Computational models of personality recognition through language.

Mark, L., & Ratliffe, K. T. (2011). Cyber worlds: new playgrounds for bullying. *Computers in the Schools, 28*(2), 92–116.

McCord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. In *Autonomic and trusted computing* (pp. 175–186). Springer.

Miller, Z., Dickinson, B., & Hu, W. (2012). *Gender prediction on Twitter using stream algorithms with N-gram character features*.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). *Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose*. arXiv preprint arXiv:1306.5204.

Myslín, M., Zhu, S.-H., Chapman, W., & Conway, M. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of Medical Internet Research, 15*(8), e174.

Nalini, K., & Sheela, L. J. (2015). Classification of Tweets using text classifier to detect cyber bullying. In *Emerging ICT for bridging the future-Proceedings of the 49th Annual convention of the Computer Society of India CSI* (Vol. 2, pp. 637–645). Springer.

Navarro, J. N., & Jasinski, J. L. (2012). Going cyber: using routine activities theory to predict cyberbullying experiences. *Sociological Spectrum, 32*(1), 81–94.

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). How old do you think I Am?; A study of language and age in Twitter. In *Proceedings of the seventh international AAAI Conference on weblogs and social media*. AAAI Press.

Nguyen, D.-P., Gravel, R., Trieschnigg, R., & Meder, T. (2013). *How old do you think I am? A study of language and age in Twitter*.

O'Keeffe, G. S., & Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. *Pediatrics, 127*(4), 800–804.

Paul, M. J., Dredze, M., & Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLoS Currents, 6*.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on search and mining user-generated contents* (pp. 37–44). ACM.

Pennacchiotti, M., & Popescu, A.-M. (2011). *A machine learning approach to Twitter user classification* (Vol. 11, pp. 281–288). ICWSM.

Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., et al. (2015). *The role of personality, age and gender in tweeting about mental illnesses*. NAACL HLT 2015 (p. 21).

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. *PLoS One, 9*(1), e86191.

Provost, F. J., & Fawcett, T. (1997). Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *KDD* (Vol. 97, pp. 43–48).

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter profiles, our selves: predicting personality with Twitter. In , *IEEE third international conference on (pp. 180–185)Privacy, security, risk and trust (PASSAT) and 2011 IEEE Third Inernational Conference on social Computing (SocialCom), 2011*. IEEE.

Räbiger, S., & Spiliopoulou, M. (2015). A framework for validating the merit of properties that predict the influence of a twitter user. *Expert Systems with Applications, 42*(5), 2824–2834.

Rangel, F., & Rosso, P. (2013). Use of language and author profiling: identification of gender and age. *Natural Language Processing and Cognitive Science, 177*.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on search and mining user-generated contents (pp. 37–44)*. ACM.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems* (pp. 532–538). Springer.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on (Vol. 2, pp. 241–244)*IEEE.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on world wide web (pp. 851–860)*. ACM.

Salmivalli, C. (2010). Bullying and the peer group: a review. *Aggression and Violent Behavior, 15*(2), 112–120.

Sampasa-Kanyinga, H., Roumeliotis, P., & Xu, H. (2014). Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren. *PLoS One, 9*(7), e102145.

Sanchez, H., & Kumar, S. (2011). *Twitter bullying detection. ser. NSDI, 12*, 15–15.

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). *Author profiling: Predicting age and gender from blogs*. Notebook for PAN at CLEF (pp. 119–124).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One, 8*(9), e73791.

Slonje, R., & Smith, P. K. (2008). Cyberbullying: another main type of bullying? *Scandinavian Journal of Psychology, 49*(2), 147–154.

Sourander, A., Klomek, A. B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., et al. (2010). Psychosocial risk factors associated with cyberbullying among adolescents: a population-based study. *Archives of General Psychiatry, 67*(7), 720–728.

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the 2015 IEEE/ACM International Conference on advances in social networks analysis and mining 2015 (pp. 280–285)*. ACM.

Talebi, M., & Kose, C. (2013). Identifying gender, age and education level by analyzing comments on Facebook. In *Signal processing and communications applications conference (SIU), 2013 21st (pp. 1–4)*. IEEE.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

Tokunaga, R. S. (2010). Following you home from school: a critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior, 26*(3), 277–287.

Vandebosch, H., & Van Cleemput, K. (2009). Cyberbullying among youngsters: profiles of bullies and victims. *New Media & Society, 11*(8), 1349–1371.

Van Royen, K., Poels, K., Daelemans, W., & Vandebosch, H. (2015). Automatic monitoring of cyberbullying on social networking sites: from technological feasibility to desirability. *Telematics and Informatics, 32*(1), 89–97.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.

Vayena, E., Salathé, M., Madoff, L. C., Brownstein, J. S., & Bourne, P. E. (2015). Ethical challenges of big data in public health. *PLoS Computational Biology, 11*(2), e1003904.

Wang, A. H. (2010a). Detecting spam bots in online social networking sites: a machine learning approach. In *Data and applications security and privacy XXIV* (pp. 335–342). Springer.

Wang, A. H. (2010b). Don't follow me: spam detection in twitter. In *Security and cryptography (SECRYPT), proceedings of the 2010 international conference on (pp. 1–10)*. IEEE.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014). Cursing in english on twitter. In *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing* (pp. 415–425). ACM.

Weir, G. R., Toolan, F., & Smeed, D. (2011). The threats of social networking: old wine in new bottles? *Information Security Technical Report, 16*(2), 38–43.

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying via social media. *Journal of School Violence, 14*(1), 11–29.

Williams, K. R., & Guerra, N. G. (2007). Prevalence and predictors of internet bullying. *Journal of Adolescent Health, 41*(6), S14–S21.

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on information and knowledge management* (pp. 1980–1984). ACM.

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012). 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). Association for Computational Linguistics.

Yang, C., Harkreader, R., Zhang, J., Shin, S., & Gu, G. (2012). Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st International Conference on world wide web* (pp. 71–80). ACM.

Yang, Y., & Pedersen, J.O. In. A comparative study on feature selection in text categorization.

Yardi, S., Romero, D., & Schoenebeck, G. (2009). Detecting spam in a twitter network. *First Monday, 15*(1).

Zhang, H. (2004a). The optimality of naive Bayes. A A, 1(2), 3.

Zhang, H. (2004b). *The optimality of naive Bayes*.

Zheng, X., Zeng, Z., Chen, Z., Yu, Y., & Rong, C. (2015). Detecting spammers on social networks. *Neurocomputing, 159*, 27–34.