Int. J. Mach. Learn. & Cyber. (2017) 8:677–689
DOI 10.1007/s13042-015-0463-1

ORIGINAL ARTICLE

Content based approach to find the credibility of user in social networks: an application of cyberbullying

Geetika Sarna[1] · M. P. S. Bhatia[1]

Received: 15 June 2015 / Accepted: 11 November 2015 / Published online: 27 November 2015

Abstract Cyberbullying is derogatory act carried out intentionally by sending or posting harmful material on social networks to cheat or tarnish anybody's image in real world. Today it has become a significant problem among teenagers and kids as they spend much time on social networking. Two types of cyberbullying have been observed in the messages posted on social network: direct cyberbullying and indirect cyberbullying. Direct cyberbullying is to send disrespectful/abusing material in the form of text, images, videos and audios to harass/torture individual directly. Indirect cyberbullying is to attack or torture individuals indirectly by doing activities like sending objectionable contents such as false rumors, lies etc. concerning them, tagging their embarrassing images, refuse to socialize with the victim. These type of activities can be viewed by large number of audience on social media. Ground breaking research is being carried out only on the identification of cyberbullying and not on its categories such as direct and indirect cyberbullying. As indirect cyberbullying is much harmful than direct cyberbullying due to the messages posted online are visible to large number of users, which may adversely impact the victim's reputation/position. So, it becomes necessary to find the solution for this problem. In this paper, we first categorize the messages into direct and indirect bullying messages and then proceed to find the solution for controlling the bullying through checking the credibility of user.

Keywords Social network · Direct cyberbullying · Indirect cyberbullying · Multigraph · Credibility

1 Introduction

Social networks are the Internet sites where people interact, sharing, discussing and disseminate knowledge/observation/information for benefit of others and possible solution of problems freely using a multimedia mix of text, pictures, videos and audio. People are using social network for their mutual benefits through information sharing but there is another side of coin also. Some of the users are misusing these social networks for the sake of their own enjoyment or showing their power by harming, harassing, insulting and abusing the other users. These types of harmful activities is called bullying.

Bullying is a behavior that usually takes place between two individuals or group. It is physically or verbally hurting another person or group. According to Kowalski et al. [1], bullying is an imbalance of power or strength. If this definition is transferred into cyberspace, and it becomes cyberbullying. Cyberbullying is "sending or posting harmful material or engaging in other forms of social aggression using the Internet or other digital technologies" [2].

Bullying is an aggressive act, still there is a difference between aggression and bullying. Bullying includes factors like intention to harm repetition in erratic behaviour and power imbalance whereas aggression is related to the behaviour that is carrying out to harm others. There are two types of aggression: Proactive and reactive aggression. Proactive aggression is to push individual in order to make him/her first in queue. So, it is not bully as it does not harm anyone. Reactive aggression is to decide socially exclude

✉ Geetika Sarna
geetika_g2002@yahoo.co.in

[1] Computer Engineering Department, Netaji Subhas Institute of Technology, Azad Hind Fauj Marg, Sector 3, Dwarka, New Delhi, India

and exert aggression through conversation and rumour spreading about the victim with the intent to harm and showing power/strength over others [3].

There are primarily six types of cyberbullying: Flaming (Online fights using messages posted online with harassing and vulgar language), harassment (insults or threats against the cyber-victim), denigration (spreading damaging rumours with intention to harm the cyber-victim's reputation), impersonation (assuming a fake identity to imitate the cyber-victim and behaving in an embarrassing or damaging way), outing and trickery (attaining and then violating the trust of the cyber-victim by publicly disclosing private and embarrassing secrets, for example via photos or videos), and exclusion (systematically excluding the cyber-victim from online activities or online groups) [4].

Figure 1 shows the classification of messages. Messages are classified into two categories: cyberbullying and non-bullying messages. Cyberbullying messages are further categorize into two categories: direct cyberbullying and indirect cyberbullying. Direct cyberbullying is to repeatedly text objectionable threatening material or abusing messages, emailing inappropriate video to individuals directly. Flaming and direct harassment come under direct cyberbullying as it is online fight between two users. Indirect cyberbullying is tagging an embarrassing photo of individual on social networking sites which can be seen by anybody, post a rude or abusive email about others to your contacts, spread rumours, lies about individuals, try to evade to socialize with the victim, bullying other people who wish to interact with the victim, associating in a poll asking people to vote on a humiliating topic, like 'who is the most irresponsible Politician in India', 'who is the ugliest celebrity in India'. Indirect harassment, denigration, outing and trickery and exclusion come under indirect cyberbullying.

The vital application of cyberbullying detection is to prevent the posting of uncalled material or harassment material on social network to preserve users' reputation. Most of the people looking for jobs online and most of the companies recruit candidates online by looking their profiles which is posted online. Any spread of negative or wrong information regarding the user may put his/her profile into risk which may threaten his/her job. Also the wrong or negative information may affect the reputation of person in family relations. His/her trustee friends may lose trust on him/her. This wrong information or rumours may also destroy the financial credibility of user in the banking sector as banks can check the reputation of customer online before approving the loan occasionally. Existing research targeted only the detection of cyberbullying and did not focus on the different categories of bullying which will be helpful in finding the credibility of user in social network. In direct bullying, we have information about the source of bullying as it is done directly but in indirect bullying, source is not known as it is done through sending rumours, post harassing/embarrassing material etc. about individual who is not in front of you. So in this case, it is difficult to find the source and also disrespectful information may be seen by large number of audience as this information spreads rapidly on the social network. Also a user whose intention is to harm any third person can harm anyone. So, it becomes essential to control such type of users. Also, existing research found the anomalous users based on the parameters related to social network structure. In this paper, we found the credibility of user based on the messages posted by the user on social network which is a new approach of doing this.

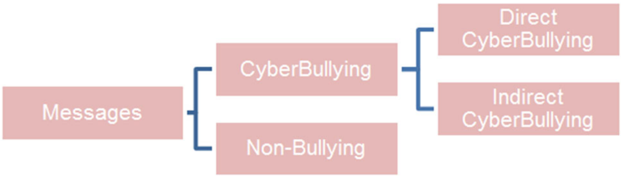

Fig. 1 Classification of messages

Proposed work can be used as a tool to identify and control the direct and indirect cyberbullying. Also it can be used to find the credibility of user. In this paper, we are concentrating on spreading of negative, harassing, embarrassing messages which are responsible for damaging the victim's reputation. After detecting such information and user involved in such type of messages, we found the credibility of users, so that the other user can alert and think while communicating. Also administrators can or give warnings to those users to make the social network more reliable which will also increase the confidence of user for using the social networks.

The rest of the paper is organized as under. We give related work in Sect. 2. In Sect. 3 we give proposed framework. In Sect. 4, we give the expected results on various real world network dataset like Twitter. In Sect. 5, we conclude the paper and outline the future course of work.

2 Literature survey

In recent study on cyberbullying detection, Dadvar et al. [5, 6], detect cyberbullying from two perspectives: the users' behavior and cross-system analysis of users' behavior. In first case, author classify the bullying and non-bullying messages based on the user's characteristics like age and gender as author assume that written language used by an individual varies with the features like age and gender, which also helpful in getting better results in cyberbullying detection. And in second case, author analyzes reaction of user against the harassing messages. Also they analyze whether user is posting vulgar messages to all users on different platform or he is targeting only single user for bullying.

In [7], author proposed a multi-criteria evaluation system (MCES) to obtain a better results in identifying the users behaviour and their characteristics. They categorize 11 features of user into 3 categories i.e. user features, content features and activity features. MCES is a framework that is used for rank the importance of features, standardize and set criterion, and combine the criteria based on the knowledge of experts to evaluate the features for a user. This technique is also applied to assign the bulliness score to the user. In [8], author make hybrid system by combining the results of MCES to machine Learning classifiers to get more accurate results.

Ptaszynski et al. [9], detect cyberbullying based on the emotion expression and emotion elements. The system they create has 2 phases: training phase and testing phase. They detected the cyber-bullying entries using support vector machine (SVM).Then harmful entries are detected and analyzed. After analysis of the results they perform one more analysis of the data, using an affect analysis system to find out how the machine learning model could be improved.

Kontostathis et al. [10], detect patterns formed by bully and their victim to detect the cyberbullying. Author mainly interested in number of bad words and density of bad words for input to the learning tool.

Authors in [11] detect cyberbullying by analyzing words that are used by cyberbullies, and the context around these words. They extended their previous work by identifying terms related to cyberbullying and generate queries using identified terms to detect cyberbullying content. There were 296 bad terms found on website www.noswearing.com and out of which 176 do not appear in the corpus and only 120 terms are left for making the queries. Four types of content query were generated. Content query 1 which is based on precision >0.5 and there are about 25 terms who meet this criteria present in this query. Content query 2 contain words that generate highest recall and there are 39 terms in this query. Content query 3 based on precision level at rank 10 means produce top results. And about 48 terms are present in this list. Content query 4 contains all the terms in bad word dictionary and ran for comparison purpose.

Dinakar et al. [12], applied binary and multiclass classifiers on a manually labelled dataset of YouTube comments. They performed two experiments. In first experiment, they used three labels (race and culture, sexuality and intelligence) individually to train binary classifier to predict that given instance belong to label or not. In second experiment, they train multiclass classifier by combining three datasets to form one dataset. The feature space is built in iterative manner. Their results showed that binary individual topic-sensitive classifiers performed well as compared to merge dataset or multiclass classifiers for the detection of textual cyberbullying.

Dinakar et al. [13], divide the problem of topic modelling into detecting features, namely, profanity and contextually relevant patterns of abuse, the use of negative language that shows profanity, as well as the employment of detailed designed to insult another person. They analysed that this method does not perform well if there is no explicit profane or negative language. Then they describe common sense reasoning model which can address this limitation. For common sense reasoning model, they developed modules like Open Mind Common Sense Knowledge Base (OMCS), Analogy base Inference Technique, The Blending Knowledge Combination Technique, The BullySpace Knowledge Base. They use statistical supervised machine learning methods [JRIP, J48 (C4.5), SVM] to detect bullying. AnalogySpace is an orthogonal transformation of the original concept and feature spaces, dot products in AnalogySpace approximate dot products in the original spaces. This fact can be used to compute similarity between features in AnalogySpace. Blending can be used to incorporate information about stereotypes to create a space more suitable for a particular application. In BullySpace Knowledge base, they build a knowledge base about commonly used stereotypes. Common sense can be used to fill the gaps between both structured and unstructured knowledge sources, or it can be designed to cover knowledge surrounding a narrow special topic.

Nahar et al. [14], proposed a novel statistical detection approach, which efficiently identifies hidden bullying. Author also presents a graph model to detect association between various users in the form of predators and victims to see the type of association between them. They used two types of feature selection methods: common feature which is mixed bullying and nonbullying features extracted using bag-of-word approach and sentiment features which are generated by applying the probabilistic latent semantic analysis (PLSA) only on bullying post. They find the predators and victim score by using the hyperlink-induced topic search (HITS) method and then find the most influential predator and victim.

Nahar et al. [15], developed augmented training technique which automatically extracts and enlarges training set from the unlabelled streaming text. This technique is suitable for real world dataset where data is very noisy, uncertain, unbalanced and labelled instances of bullying data are not available always. Feature which the author used to train the classifier are: keyword based features, influence of malevolence within messages, presence of pronouns, degree of user's emotions, capital letters which indicate shouting, metadata of messages, user's age and gender. They used Fuzzy SVM (FSVM) rather than SVM. They incorporate membership function to FSVM. They used kernel based Fuzzy C-Means (K-FCM) clustering algorithm to generate membership values for our fuzzy classifier model to handle noise and uncertainties.

Serra et al. [16] focused on the usage of social networking sites through mobiles technologies like cell phones rather than the static configuration. Most of the youth involved in online activities through the cell phones which also increase the chances of engaging in the criminal activities. They proposed a solution for the detection of cyberbullying on mobiles technologies. They tried to find risked profile of user on the basis of three parameters: age, time spent online, types of online access and associated risks and then identify the threat based on these profiles and suggest protection accordingly.

Li et al. [17], generated contact network to detect the friendship. This study focuses on the detection of indirect attack on a human, e.g., isolating a victim by ignoring the victim's messages. There are two phases of proposed framework: construction phase and expansion phase. The construction phase is used to create undirected graph from the data collected in dataset and then expansion phase is used to add missing links based on the hop count and similarity method to create the contact network. False negative is the main problem of this study.

Galán-García et al. [18], detected of fake profiles on twitter by analyzing the content generated by both the troll profiles as well as real user profiles. Feature used are tweets which gives the writing style of user, time of publication, language and geoposition which shows the behavior of the user and finally the tweet client like devices from where user tweet.

Michael et al. [19], identified fake user in Facebook. They developed the software with three protection layers. First layer is responsible for identifying those friends who might pose threat and then restrict them to access user's personal information. Second layer is an expanded version of Facebook's basic privacy settings. The third layer is responsible for alerting the user about application installed on Facebook profile that has access to their private information. For finding the connection strength between user u and his/her friend v, they extract features like Are Family(u,v), Common Chat messages(u,v), Common friend(u,v), Common Groups Number(u,v), Common Post Number(u,v), Tagged Photos Number(u,v), Tagged videos Number(u,v), Friend's Number(u) and Friend's Number(v). Connection Strength between user u and his/her friend v is calculated as:

$$\begin{aligned}\text{Connection Strength }(\text{u,v}) = {} & 1000 \times \text{ Are Family}(\text{u,v}) \\ & + \text{ Common Chat messages}(\text{u,v}) + \text{Common friend}(\text{u,v}) \\ & + 2 \times \text{ Common Groups Number}(\text{u,v}) \\ & + 2 \times \text{Common Post Number}(\text{u,v}) \\ & + 2 \times \text{Tagged Photos Number}(\text{u,v}) \\ & + 2 \times \text{Tagged videos Number}(\text{u,v})\end{aligned}$$

Also they analyze and monitored the privacy settings on Facebook by extracting the features like Installed application number, Default privacy settings, Look up, Share address, Send messages, Receive Friend requests, Tag suggestions and view birthday.

They defined four type of link set: All unrestricted links set, Recommended Unrestricted link set, Recommended Restricted link set, Alphabetically Restricted link set for restricting the user. Based upon these link sets, authors defined three datasets: Fake Profile Dataset, Friends Restriction Dataset and All Links Dataset. For each link in above dataset, author found the following features: Chat Messages Ratio(u,v), Common Group Ratio(u,v), Common Post Ratio(u,v), Common Photo Ratio(u,v), Common Video Ratio(u,v), Is Friend Profile Private(u,v), Jaccard's Coefficient(u,v). Using the above dataset and features defined, they classify the data using WEKA's C4.5, IBK, Naïve Byes, Bagging, AdaboostM1, and Random forest classifiers.

Simon et al. [20], detected imposters or fake profiles on Facebook using decision tree classifiers. Features used are Age, Gender, College Degree, Avatar photo, Personal information in the profile, Authentic pictures, advertisements, Profile Completeness, Number Of friends, length Of membership, gender of Majority of mutual friends, comments on their posts. They applied five different algorithm of decision tree implemented in WEKA: J48, REPTree, Random tree, ADTree, Functional tree (FT). The accuracy varies from 70.3 to 92.1 % depending on the type of algorithm used.

Charles et al. [21] detected suspicious profiles on social platforms on smartphones. They used the dataset of Twitter. They identified two core indicators: activity of profile and visibility of profile. Activity of profile is defined as the number of actions it performs during the time period T and Visibility of profile is defined as the amount of techniques it performs to increase the audience during the time period T. Visibility is measured based on the number of keywords present in the message and references included in messages. Then they considered three features based on the core indicators activity and visibility. First feature is the balance between the number of messages sent with the visibility. Second feature is the energy that is consumed by the profile to increase the level of activity and visibility of

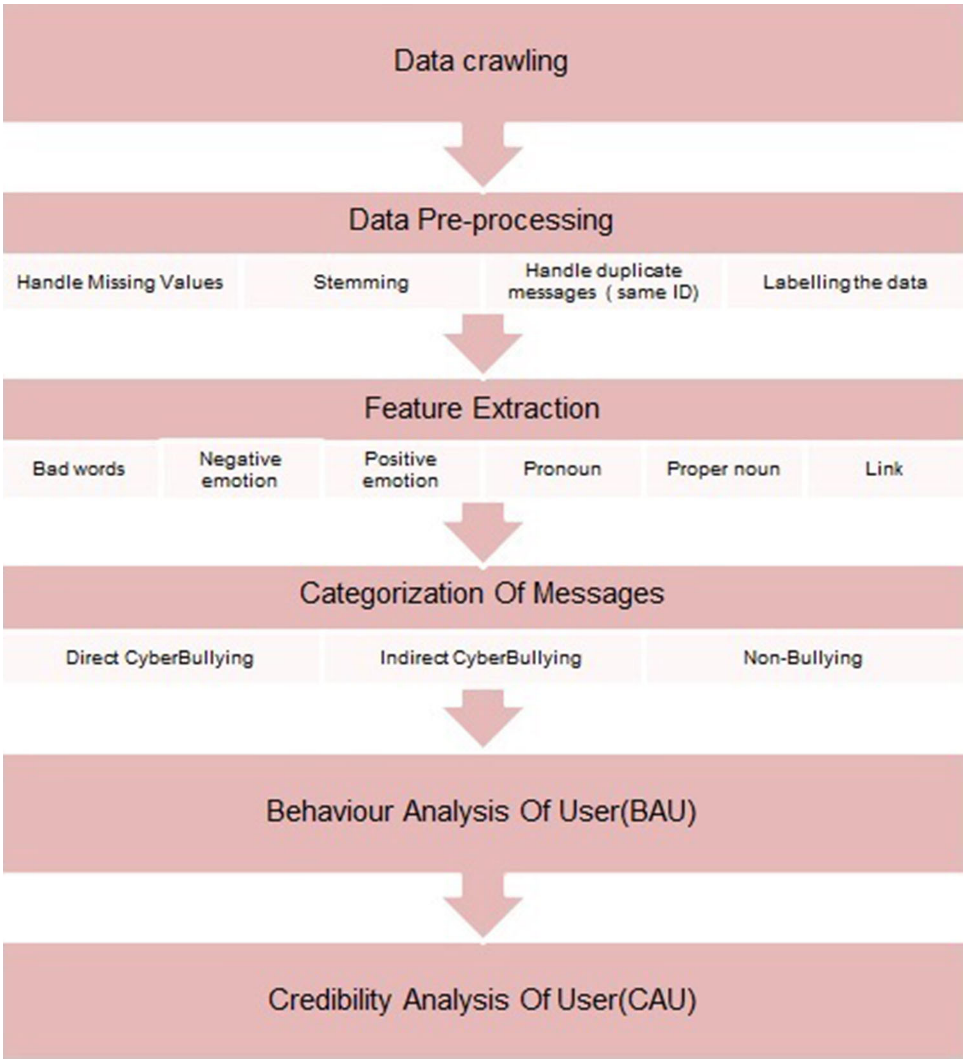

Fig. 2 Proposed Framework for finding the credibility of user

profile. Third indicator is the anomaly score of activity–visibility pair. The chances of the suspicious profiles of having the unusual activity–visibility pair are more than the normal profiles. They used the naïve bayesian classifier to classify the suspicious and non-suspicious profiles. This method showed high performance within a time frame of three to 5 days only.

Conti et al. [22], identified fake profiles in Online Social Network base on the social network graphs. They used the dataset of Facebook. They considered three features: Evolution overtime of the number of online social network (OSN) friend, Real social interaction, Evolution overtime of the structure of OSN. First feature is concerned with detecting whether the profile under investigation matches with the profile of number of friends added over time. Second feature is to identifying the adversary by considering the high frequency in-person contact friend to the fake profile. Third feature is related to observing whether the network structure of the profile under investigation is anomalous when compared to the social network of total population. This method is efficient for the dynamic social network.

Liu et al. [23], predict trust between users in online communities. They used the dataset of Epinion, the largest review community. They considered two factors: user factor and interaction factor and extract features based on these factors. They identified three groups of user: review writer, review rater and review commenter. For the entire three groups above, they are further divided into two subgroups: distribution factors and count-based factors. Distribution factors can be calculated by the statistics metrics such as average and standard deviation, while count-based factors are related to counting a particular set of objects. As each interaction involves two users, the interaction factors are categorized into three groups: trustor, trustee, and a connection specific temporal factor (the time difference between two user's actions). They used decision tree, Naive Bayes (NB), logistic regression, and

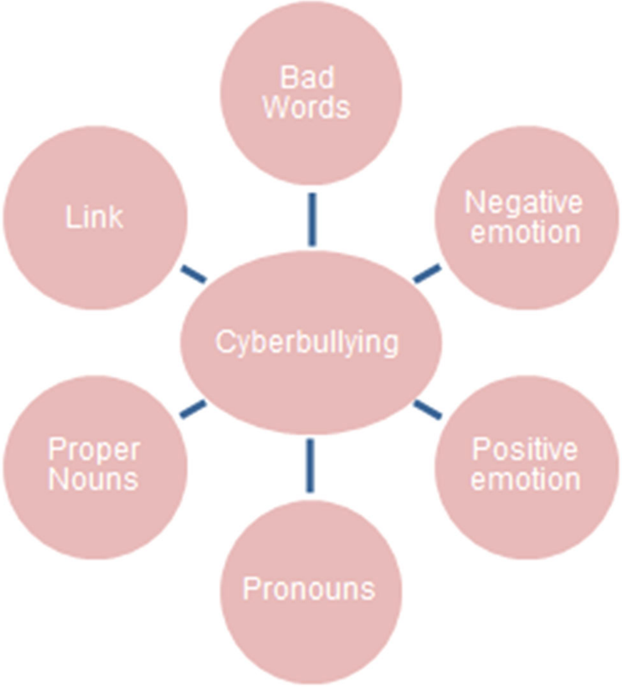

Fig. 3 Problem based features responsible for the detection of categories of Cyberbullying

SVMs with linear kernel and RBF kernel for classification of trusted and untrusted user.

3 Proposed framework

As we already mentioned that the detection of direct cyberbullying only is not sufficient to prevent the social network from bullies. There are a couple of users who are responsible for indirect cyberbullying also which is more harmful. So we need to consider both types of cyberbullying for finding the credibility of user. Figure 2 shows the proposed framework for finding the credibility of user.

Proposed framework consists of following modules:

- Data crawling
- Data pre-processing
- Feature extraction
- Categorization of messages
- Behaviour analysis of user (BAU)
- Credibility analysis of user (CAU)

3.1 Data crawling

In this module, we captured data from Twitter by using our customized crawler written in Python. Data we got from twitter contain direct cyberbullying, indirect cyberbullying, implicit harassment and sarcasm but we focused only on direct and indirect cyberbullying.

3.2 Data pre-processing

Data captured from twitter contains many missing fields, duplicate tweets. Pre-processing of the dataset involve following steps:

1. Missing fields are replaced by NULL.
2. Stemming.
3. There are large numbers of duplicate tweets in dataset. We deleted these duplicated tweets based on TweetID.
4. Data we got is also not labelled. Thus we manually labelled the data. The messages we extracted from Twitter are divided into three categories: direct cyberbullying, indirect cyberbullying and no bullying. Now the data of dataset become multiclass.

3.3 Feature extraction

As we have already mentioned above that the messages we got are having direct cyberbullying, indirect cyberbullying and no bullying. Figure 3 represents problem based features responsible for the categorization of cyberbullying messages which are described below:

- *Bad words*: The primary use of this feature is to detect message containing bad words from the large dataset. This feature behaves like a filter as it filters out the normal messages (messages without bad words) from large dataset and messages with bad words are left behind. If there is large number of bad words in the message, there are more chances of being it a bullying message.
- *Negative emotions*: Negative emotion is the strong indicator of bullying behaviour as it puts negative impact on the victim. For example: negative emotional words like hate, ugly, cheap etc. put more depression on victim. This feature can give more accurate results in the classification of messages.
- *Positive emotions*: It is not necessary that messages containing bad words are always bullying messages. Positive emotions feature is helpful in detecting the non-bullying behaviour in the messages even though the messages contain bad words. For example: “You are my darling bitch”. As this message contains bad word “bitch” but it also contain positive emotional word “darling” which indicates that this messages is not a bullying message. Bad words should not be the only criteria for capturing the bullying messages and bullies. Using such type of words might be the writing style of users.

- *Pronouns*: This feature gives meaning to the bullying messages means to which user the message is associated with. There are three types of pronouns. First person pronoun e.g. I, We. Second person pronoun e.g. you. And Third person pronoun e.g. he, she, her, his etc. These types of pronoun are helpful in detecting the direct and indirect bullying behaviour of the messages. For example:
    - *First example*: "I hate you.". This message is the combination of first person pronoun, words showing negative emotions and second person pronoun which is helpful in capturing the direct bullying.
    - *Second Example: "*You are a Kaffir**".** This message is the combination of second person pronoun with bad words which is helpful in capturing direct bullying.
    - *Third example*: **"**She is a bitch." This message is the combination of third person pronoun or proper noun with bad word which is helpful in capturing indirect bullying.
- *Proper nouns*: This feature gives some name which is helpful in detecting the indirect cyberbullying.
- *Links*: This feature is helpful in finding the severity of bullying in messages. If messages contain bad words, links, pronouns, it means messages is very harmful as it may contain embarrassing information inside that link regarding the victim and would be visible to large audience. This feature is also helpful in getting the accurate results.

3.4 Categorization of messages

In this module, we used four famous machine learning methods, which need pre-labelled training data for automatic learning: a Naive Bayes classifier, K Nearest Neighbour (KNN), a classifier based on Decision trees and Support Vector Machines (SVM). We applied these models for multiclass classification i.e. 3 class classification. As binary classification would not perform well because it had to follow two steps:

1. Classification of messages into bullying and non-bullying. As we have already discussed in Sect. 3.3 that it is not always necessary that messages containing bad words are bullying messages. So, we need to separate those messages from the bullying messages to get the most accurate results. This is the reason we added feature of positive emotions which help us to filter out non-bullying messages.
2. Classification of bullying messages into direct bullying and indirect bullying.

For the execution of the above two steps, binary classifier would take more time than multiclass classifier. So, this classifier is not the beneficial criteria for doing the classification as we have large number of messages. In comparison to this, multiclass classifier classifies the messages into three classes in one step and is faster than binary classifier.

The implementation available in MATLAB was used for classification of messages. Machine learning models classify the messages based on the following parameters:

1. The number of bad words in the messages like bitch, nigger/nigga, cheater, kaffir, black etc.
2. The number of words showing negative emotions like dislike, disgust, scary, shut up, idiot, angry, kill, throw etc. used for detecting the severe bullying messages.
3. The numbers of words showing positive emotions like love, like, loyalty, honestly, sorry, happy, good etc. are used for detecting the non-bullying messages which sometimes contain bad words also. For example: I am happy to see u bitch.
4. Combination of first person pronoun, words showing negative emotions and second person pronoun to capture direct bullying. For example: I hate you.
5. Combination of second person pronoun with bad words to capture direct bullying. For example: You are a Kaffir.
6. Combination of first person pronoun, words showing negative emotions and third person pronoun or proper noun to capture indirect bullying. For example: I hate Mr. Bose.
7. Combination of third person pronoun or proper noun with bad word to capture the indirect bullying. For example: She is a bitch.
8. Combination of link, bad words and pronouns is also used to capture bullying. If the message contains only link and bad words, then messages are non-bullying message. If it contains pronouns also then the message is considered as bullying messages and its category depends upon the use of pronouns.

It becomes cumbersome to detect exact bullying messages. It is examined that detection of exact bullying needs more features along with bad words and negative emotional words. So we added one more feature positive emotional words like happy, darling etc. to detect the non-bullying messages. In this way, we isolated non-bullying messages from total messages. After that, we divided rest of the messages into two categories direct bullying and indirect bullying based on the pronouns and proper nouns. Second person pronoun is responsible for direct bullying and third person pronouns and proper nouns are responsible for indirect bullying. Link is also very important for categorization of messages. If messages include only link and few bad words, it indicates non-bullying messages. If messages contains pronoun also then we categorize the messages into other two categories depending upon the type of pronoun it contains.

3.5 Behaviour analysis of user (BAU)

In this module, we took the help of multigraph which is helpful in representing the dynamic network by creating the parallel edges between nodes. A multigraph is a representation of a set of nodes where some pairs of nodes are connected by one or number of edges. In this step, directed multigraph is created. Nodes are sender and receiver of the messages and edge exists if message is sent from sender to receiver. Also the sequence number is assigned to the edges which indicate the order followed by message and time of that message indicating the dynamic nature of social network. With the time feature, we can see the number of messages sent per unit time which is helpful in observing the repeated behavior of user. Here Directed multigraph G is an represented by quadruple

$$\mathrm{G} = (\mathrm{S},\ \mathrm{R}, \mathrm{E}, \mathrm{W})$$

where S, sender of the message; R, receiver of the message; E, directed edge between Sender and Receiver if message sent from sender to receiver; W, weight of the edge represented by sequence number of edge followed by the message, time and type of bullying.

From this multigraph, we got total number of messages sent from sender to receiver with sequence number, category, time which is helpful in finding the behaviour of user. This information about the behaviour is helpful in controlling the user from doing the unwanted activities on social network. We divide the behavior of user into two categories:

- Normal behaviour
- Abnormal Behaviour
  - Direct Bullying
  - Indirect Bullying

These categories of behaviour depend upon following factors:

1. Total number of non-bullying messages sent by user x for time t NB(x,t)indicates the normal behaviour.
2. Total number of direct bullying DB(x, t) and indirect bullying messages IDB(x,t) sent by user x for time t indicates the abnormal behaviour.

From the number of messages, we found the probability of normal and abnormal behaviour. We measured abnormal behaviour in terms of direct and indirect bullying.

Probability of Normal Behaviour of user x for time t is given by:

$$P(Nor(x,t)) = \frac{NB(x,t)}{(NB(x,t) + DB(x,t) + IDB(x,t))} \tag{1}$$

Probability of Direct Bullying of user x for time t is given by:

$$P(DBull(x,t)) = \frac{DB(x,t)}{(NB(x,t) + DB(x,t) + IDB(x,t))} \tag{2}$$

Probability of Indirect bullying of user x for time t is given by:

$$P(InBull(x,t)) = \frac{IDB(x,t)}{(NB(x,t) + DB(x,t) + IDB(x,t))} \tag{3}$$

Equations (1–3) are used to find the probability of the normal or abnormal Behaviour of the user.

3.6 Credibility analysis of user (CAU)

CAU is basically a rule based system which depends upon the output of BAU module. Indirect bullying is more harmful than direct bullying for the victims as sender can post any type of harassing material about anyone which may harm his/her reputation. Everybody on the social network can view those messages about the victim which may affect his/her profile. Job Seekers, recruiters can see the information posted about the victim on social network. So, we give higher priority to indirect bullying rather than direct bullying. Now Credibility of user is calculated by using the information passed by BAU module and by using below 12 rules:

1. If P (Nor(x,t)) $\geq$ 0.5 and P(DBull(x,t)) $\leq$ 0.4 and P(InBull(x,t)) $\leq$ 0.2) then credibility = Normal user.
2. If P (Nor(x,t)) $\geq$ 0.5 and P (DBull(x,t)) $\leq$ 0.4 and P (.2 < P (InBull(x,t)) $\leq$ .5) then credibility = Harmful/selfish
3. If P (Nor(x,t)) $\geq$ 0.5 and P (DBull(x,t) > 0.4 and P (InBull(x,t)) $\leq$ 0.2) then credibility = Normal user
4. If P (Nor(x,t)) < 0.5 and P (DBull(x,t)) $\leq$ 0.4 and P (InBull(x,t)) > 0.2) then credibility = Very harmful/Trustless
5. If P (Nor(x,t)) < 0.5 and P (DBull(x,t)) > 0.4 and P (InBull(x,t)) > 0.5) then credibility = Very harmful/Trustless
6. If P (Nor(x,t)) < 0.5 and P (DBull(x,t)) > .4 and (.2 < P (InBull(x,t)) $\leq$ .5)then credibility = Harmful/selfish.
7. If P (Nor(x,t)) < 0.5 and P (DBull(x,t)) $\leq$ 0.4 and P (InBull(x,t)) $\leq$ 0.2) then credibility = Harmful/selfish.

8. If P (Nor(x,t)) < 0.5 and ($0.4 <$ P (DBull(x,t)) ≤ 0.8) and P (InBull(x,t)) ≤ 0.2) then credibility = Harmful/selfish.
9. If P(Nor(x,t)) < 0.5 and ($0.4 <$ P(DBull(x,t)) ≤ 0.8) and ($0.2 <$ P(InBull(x,t)) ≤ 0.5) then credibility = Harmful/Selfish
10. If P (Nor(x,t)) < 0.5 and (P (DBull(x,t)) > 0.8) and (P (InBull(x,t)) > 0.2) then credibility = very Harmful/trustless
11. If P (Nor(x,t)) < 0.5 and (P (DBull(x,t)) > 0.8) and (P (InBull(x,t)) < 0.2) then credibility = Harmful/selfish
12. If P (Nor(x,t)) < 0.5 and (P (DBull(x,t)) < 0.8) and (P (InBull(x,t)) > 0.5) then credibility = very Harmful/trustless

Here, credibility of user depends upon the his/her probability of normal and abnormal behaviour. As the network is dynamic and number of messages are growing with time, threshold value depends upon the probability rather than number of messages. We assumed that the indirect bullying is very harmful so we set the threshold value for indirect bullying as 0.2 means if P(InBull(x,t)) goes beyond 0.2, risk starts. And if the threshold goes beyond 0.5 then user become more harmful. Similarly, we set the threshold value for direct bullying. As it is less harmful than indirect bullying, we set 0.4 as threshold value means if P(DBull(x,t)) goes beyond 0.4, risk starts. And if the threshold goes beyond 0.8 then user become more harmful. Also we set the threshold for normal behaviour as 0.5 means if P(Nor(x,t)) goes beyond 0.5, no risk else risk starts.

4 Results

Initially, we extracted the messages from twitter using crawler based on some keywords i.e. bad words. In this way, we get a small portion of large dataset and that portion contains only bad words. As we are considering the bullying messages, we need to concentrate on bad words first.

Next, we detected nouns and pronouns from messages using Stanford Postagger software. A Part-Of-Speech (POS) Tagger is a software that accept text in some language and assigns parts of speech tags to each word (token), such as noun, pronoun, verb, adjective, etc. This software consists of the log-linear part-of-speech taggers. Following are the Part-Of-Speech tag which we have used to find the nouns and pronouns to make the feature set.

- NNP Proper noun, singular

  e.g. Shannon, A.K.C., Meltex, Liverpool

- NNPS Proper noun, plural

  e.g. Americans, Americas, Amharas

- PRP Personal pronoun

  e.g. Hers, herself, him, himself, it, itself, me, myself, one, oneself, ours, ourselves, ownself, self, she, thee, theirs, them, themselves, they, thou, thy, us.

- PRP$ Possessive pronoun

  e.g. her, his, mine, my, our, ours, their, thy, your.

Stanford Postagger is not beneficial for large dataset as it become very slow during parsing and assigning the tags to each word. So, we have taken a small portion of large dataset containing bad words and Stanford Postagger works well for that. Also this software is implemented in Java, so is platform independent which is the main advantage of this software.

As the proposed method is novel approach to find the credibility of user, we applied exiting classifiers to compare the performance. We applied four machine learning classifiers KNN, Decision tree, Naïve Bayes Classifier, SVM on the dataset. We divide the dataset into two part: 50 % of dataset is used for training the classifier and 50 % of dataset is used for testing the classifier. Then we found the three classification errors: Resubstitution error, Cross validation error and Test error, Kappa Statistics and classifier performance in terms of Precision, Recall, F_Score. After that, we showed the comparison of three classifiers in terms of Receiver Operating Characteristic (ROC) curve.

Table 1 shows Classification error and Kappa Statistics for Decision Tree, Naïve Bayes Classifier, SVM, KNN.

Table 2 shows the classifier performance for different category of messages. These results showed that KNN is better than other classifiers as it has better results regarding classification error and kappa statistics.

Figure 4 shows the comparison of KNN, Decision tree, SVM, Naïve Bayes Classifier through ROC curve. Although it shows little bit better results for decision tree but we get better results for classification error, Precision, Recall and F_Score in case of KNN.

A ROC curve is representing True Positive Rate (TPR) against False Positive Rate (FPR). As there are more chances that number of negative examples exceeds the number of positives examples, a large modification in the number of false positives can lead to a small modification in the false positive rate used in ROC analysis. And Precision, on the other hand, compare false positives to true positives rather than true negatives and shows strong effect on the algorithm's performance [24]. So, we decided to use KNN for predicting the data.

After classification we got 11 % messages are having indirect cyberbullying, 18 % messages are having direct

Table 1 Classification error and kappa statistics for decision tree, Naïve Bayes classifier, SVM, KNN

	Decision tree (%)	Naïve Bayes classifier (%)	SVM (%)	KNN (%)
Resubstitution error	17.50	28.25	27.75	13.25
Cross validation error	29.75	31.31	28.25	28.50
Test error	28.38	23.91	22.38	29.17
Kappa statistics	62.34	39.82	35.55	72.32

Table 2 Classifier performance for different category of messages

Classifier	Measure	Non-bullying	Direct bullying	Indirect bullying
Decision tree	Precision	.8522	.7627	.7400
	Recall	.9185	.7031	.5606
	F_Score	.8841	.7317	.6379
Naïve Bayes classifier	Precision	.8038	.4955	.5000
	Recall	.7889	.8594	.1818
	F_Score	.7963	.6286	.2667
SVM	Precision	.7610	.5696	.6667
	Recall	.8963	.7031	.0303
	F_Score	.8231	.6294	.0580
KNN	Precision	.8929	.8667	.7500
	Recall	.9259	.8125	.6818
	F_Score	.9091	.8387	.7143

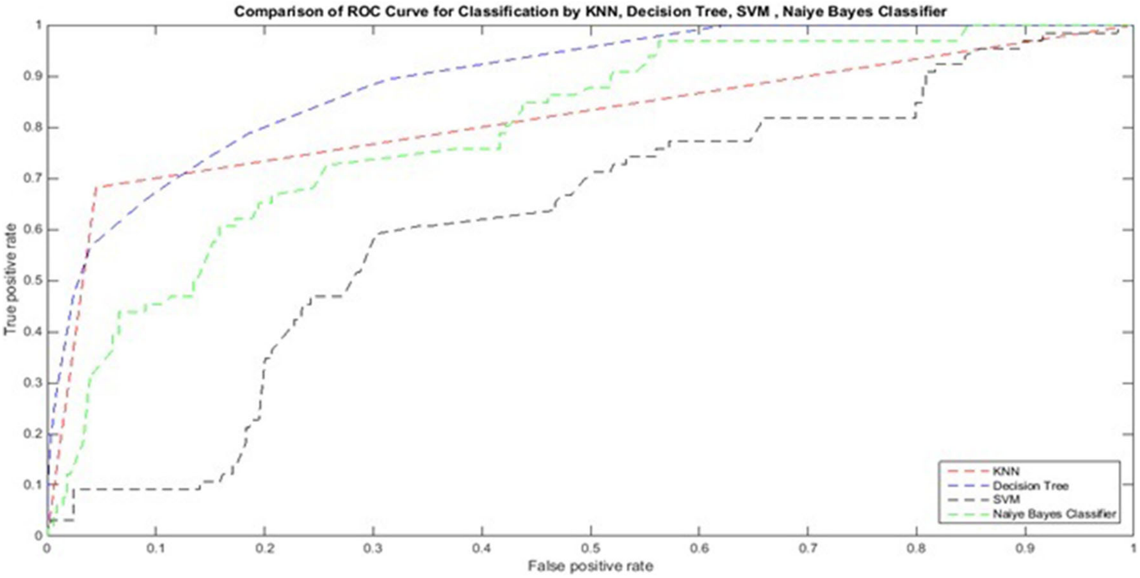

Fig. 4 Comparison Of ROC Curve for KNN, Decision tree, SVM, Naïve Bayes Classifier

cyberbullying and 71 % messages are non-bullying messages. So, it is analysed from these results that the less number of users are involved in indirect cyberbullying and direct bullying. Indirect cyberbullying is harmful as it plays with the career or life of user, it is very important to pay significant attention to control this although the percentage of indirect bullying messages is less. Also we analysed that the messages contain bullying words are not bullying messages always. We found that large number of users using bad words or bullying words to express their thoughts is responsible for non-bullying messages as it is the habit of few users to use these types of words or it is the writing style of user. So, we can’t identify that the message is bullying or not only on the basis of few bad words or negative emotions. We need lot more feature to distinguish between bullying and non-bullying messages.

Table 3 Information of multigraph including details of sender, receiver and messages

S. No.	Sender	Receiver	No. of messages	List of messages
1	BlackSpeedRacer	DannyFrio	2	{0: {‘weight’: “RT @DannyFrio: MONKEY NIGGER COON SAMBO PORCH MONKEY TAR BABY. YES, I’M ALL OF THAT BUT YOUR WHITE DAUGHTER LOVES IT SIRTue Jan 20 07:29:43 +0000 2015Direct Bullying”}, 1: {‘weight’: “RT @DannyFrio: calling me a nigger isn’t gonna stop me from fucking your beautiful white daughter sir.Tue Jan 20 07:30:54 +0000 2015Direct Bullying”}}
2	shamelessShely	eagle1776n	3	{0: {‘weight’: ‘@eagle1776n @KombuchasmithS Calling a white person a “nigger” doesn\'t not dehumanize black people.Tue Jan 20 07:23:28 +0000 2015No Bullying’}, 1: {‘weight’: ‘@eagle1776n @therightswrong @DavidIFisher You dehumanize women by using the word “whore” in the same way you dehumanize blacks with “nigger”Tue Jan 20 07:22:56 +0000 2015Direct Bullying’}, 2: {‘weight’: ‘@eagle1776n @HollyRFisher @TedHaggard @therightswrong @DavidIFisher Jesse Jackson is an ADULTEROUS NIGGER WHORE? Keep dehumanizing people…Tue Jan 20 07:20:58 +0000 2015Indirect Bullying’}}
		RedScareBot	1	{0: {‘weight’: ‘@RedScareBot @eagle1776n @KombuchasmithS @HollyRFisher Keep dehumanizing people with words like “whore” and “nigger” and watch the deathtollTue Jan 20 07:29:35 +0000 2015No Bullying’}}
		WILLYGinPDX	2	{0: {‘weight’: ‘@WILLYGinPDX @HollyRFisher So use the word “nigger” against people of all races and then you can call black people “nigger” as you please?Tue Jan 20 07:42:33 +0000 2015No Bullying’}, 1: {‘weight’: ‘@WILLYGinPDX @HollyRFisher It\'s not apples and oranges. It\'s apples and apples. “Whore” is no better than “nigger”Tue Jan 20 07:47:18 +0000 2015No Bullying’}}
3	aardnasac	NoHoesNick	7	{0: {‘weight’: ‘RT @NoHoesNick: 10. This counter lady was some fat ass girl who was lookin at me like “Awhh this little kid cute AF for this.” BITCH YOU LE?Tue Jan 20 07:30:02 +0000 2015Indirect Bullying’}, 1: {‘weight’: ‘RT @NoHoesNick: 9. I went into Party city, dropped a cold hard $20 on the counter. Looked that fat bitch Maria dead in her eye and told her?Tue Jan 20 07:29:59 +0000 2015Indirect Bullying’}, 2: {‘weight’: ‘RT @NoHoesNick: 8. I was in the car smiling happy AF like “OMG I\'m finally getting this girl.” NAHHH NIGGA YOU GONNA GET THIS FUCKIN CURVE ?Tue Jan 20 07:29:57 +0000 2015No Bullying’}, 3: {‘weight’: “RT @NoHoesNick: 4. She was ?? AF. I’m talkin ass fatter than Precious and head game was A1 at least what I heard cuz I’m a little bitch and ?Tue Jan 20 07:29:47 +0000 2015No Bullying”}, 4: {‘weight’: “RT @NoHoesNick: 20. So I’m with the homies the WHOLE game tryna play cool cuz fuck these hoes right? I’m chillin watching the game not real?Tue Jan 20 07:32:09 +0000 2015No Bullying”}, 5: {‘weight’: ‘RT @NoHoesNick: 25. THE NIGHT BEFORE THE DANCE. I was a little bitch ass heartbroken freshman. I bought my suit, the mum, her corsage, ever?Tue Jan 20 07:32:51 +0000 2015Indirect Bullying’}, 6: {‘weight’: ‘RT @NoHoesNick: 23. Man that night she went to go teepee with everyone and I guess she found a new nigga there because I got made a bitch t?Tue Jan 20 07:32:35 +0000 2015No Bullying’}}
		FaceItYouBASIC	2	{0: {‘weight’: ‘@FaceItYouBASIC oh imbitchTue Jan 20 07:35:01 +0000 2015No Bullying’}, 1: {‘weight’: “ @FaceItYouBASIC so i’m a bitch rightTue Jan 20 07:41:36 +0000 2015No Bullying”}}

Table 3 continued

S. No.	Sender	Receiver	No. of messages	List of messages
4	tKOs_way	Khleopatra_Lynn	1	{0: {‘weight’: “RT @Khleopatra_Lynn: ???? RT?@Bizzy_Bdy: ?? Jesus ?? RT @tKOs_way: I’m not a cheater but I will fuck your friends to get to your feelings?Tue Jan 20 07:25:18 +0000 2015Direct Bullying”}}
		yea_im_JAMAICAN	1	{0: {‘weight’: “RT @yea_im_JAMAICAN: ?????? RT @tKOs_way: I’ll break a bitch heart on Valentine’s Day, it happened beforeTue Jan 20 07:31:11 +0000 2015Indirect Bullying”}}
		imjust_alex	1	{0: {‘weight’: ‘RT @imjust_alex: ???? “@tKOs_way: I\’ll break a bitch heart on Valentine\’s Day, it happened before”Tue Jan 20 07:36:09 +0000 2015Indirect Bullying’}}
		Bizzy_Bdy	5	{0: {‘weight’: “I’m not a cheater but I will fuck your friends to get to your feelingsTue Jan 20 07:10:35 +0000 2015Direct Bullying”}, 1: {‘weight’: ‘I care less about how a bitch feel when they care nothing how I feelTue Jan 20 07:29:56 +0000 2015Indirect Bullying’}, 2: {‘weight’: “ @Bizzy_BdyThat bitch not beautiful and you’re not eitherTue Jan 20 07:30:20 +0000 2015Direct Bullying”}, 3: {‘weight’: “Lil Ko don’t give a fuckTue Jan 20 07:36:59 +0000 2015Indirect Bullying”}, 4: {‘weight’: ‘@Bizzy_BdyShe pull up and fuck me but she connivingTue Jan 20 07:38:27 +0000 2015No Bullying’},

Table 3 shows the result of multigraph which we created for the dataset after classification. Multigraph contains sender node which is represented by the Sender field, receiver node which is represented by the Receiver field, Edge from sender to receiver. Each edge has weight represented by sequence number followed by message, time and type of bullying. From the last column of table i.e. List of messages, we can find the number of indirect, direct and non-bullying messages sent from a user for particular time.

Example 1 In first case, 2 messages are sent from BlackSpeedRacer to DannyFrio in one day and both the messages are come under direct bullying.

Here,

$\mathrm{NB} = 0$
$\mathrm{DB} = 2$
$\mathrm{IDB} = 0$

Here (using Eqs. 1, 2, 3),

$$P(Nor(BlackSpeedRacer, 1)) = 0$$

$$P(DBull(BlackSpeedRacer, 1)) = 1$$

$$P(InBull(BlackSpeedRacer, 1)) = 0$$

Credibility analysis module shows that the credibility of user is harmful/selfish.

Example 2 In third case, 9 messages are sent from aardnasac to (NoHoesNick, FaceItYouBASIC).

Here,

$\mathrm{NB} = 6$
$\mathrm{DB} = 0$
$\mathrm{IDB} = 3$

Here (using Eqs. 1, 2, 3),

$$P(Nor(aardnasac, 1)) = .67$$

$$P(DBull(aardnasac, 1)) = 0$$

$$P(InBull(aardnasac, 1)) = .33$$

Credibility analysis module shows that the credibility of user is harmful/selfish.

5 Conclusion and future work

Cyberbullying is criminal offence as it harms the reputation of victim. So, the detection of cyberbullying messages is essential. In this paper, we divided the messages into Direct bullying, Indirect Bullying and non-bullying categories in order to detect the credibility of user. Pronouns and words with negative emotions features are appropriate in finding the bullying behaviour and words with positive emotions are extremely helpful in finding the non-bullying messages. We applied and compared machine learning classifiers on multiclass data for the categorization of messages. Then we use multigraph in order to get information of multiple messages sent by the user as the network is dynamic. Then we detected behaviour of the user based on the type of messages sent by user and based upon that information we found the credibility of user.

In future work, we aspire to predict the behaviour of user based upon the previous information using mathematical model like Markov chain model. Also, the key target is to extract new bad words from the messages and keep our dictionary up to date.

This paper focused only on finding the credibility of user based on the category of bullying he/she does. It does not reveal any status of relationship between the users. So, we intend to find the status of relationships between users using Markov random fields which will be helpful in making more accurate conclusion for someone.

References

1. Kowalski RM, Limber SP, Agatston PW (2008) Cyber-bullying bullying in the digital age, 1st edn. Blackwell, Malden
2. Willard NE (2007) Cyberbullying and cyberthreats: responding to the challenge of online social aggression, threats, and distress. Research Press, Champaign
3. Law DM, Shapka JD, Domene JF, Gagné MH (2012) Are cyberbullies really bullies? An investigation of reactive and proactive online aggression. J Comput Hum Behav 28:664–672
4. Pieschl S, Porsch T, Kahl T, Klockenbusch R (2013) Relevant dimensions of cyberbullying—results from two experimental studies. J Appl Dev Psychol 34:241–252
5. Dadvar M, Ordelman R, Jong FD, Trieschnigg D (2012) Towards user modelling in the combat against cyberbullying. Paper presented at 17th international conference on applications of natural language to information systems. Springer, Netherlands, pp 277–283
6. Dadvar M, Jong FD, Ordelman R, Trieschnigg D (2012) Improved cyberbullying detection using gender information. In: Proceedings of the Twelfth Dutch-Belgian information retrieval workshop (DIR2012), Ghent, Belgium, pp 23–26
7. Dadvar M, Trieschnigg D, Jong FD (2013) Expert knowledge for automatic detection of bullies in social networks. Paper presented at 25th Benelux conference on artificial intelligence, BNAIC, Delft, Netherlands, pp 57–64
8. Dadvar M, Trieschnigg D, Jong FD (2014) Experts and machines against bullies: a hybrid approach to detect cyberbullies. Paper presented at Canadian conference on AI, Springer, Switzerland, pp 275–281
9. Ptaszynski M, Dybala P, Matsuba T, Masui F, Rzepka R, Araki K (2010) Machine learning and affect analysis against cyberbullying. In: Proceedings of the linguistic and cognitive approaches to dialog agents symposium. De Montfort University, Leicester, UK
10. Reynolds K, Kontostathis A, Edward L (2011) Using machine learning to detect cyberbullying. Paper presented at 10th IEEE international conference on machine learning and applications vol 2, pp 241–244
11. Reynolds K, Kontostathis A, Garron A, Edward L (2013) Detecting cyberbullying: query terms and techniques. In: Proceedings of 5th annual science conference. ACM, New York, pp 195–204
12. Dinakar K, Reichart R, Lieberman H (2011) Modeling the detection of textual cyberbullying. Paper presented at fifth international AAAI conference on weblogs and social media 2011
13. Dinakar K, Jones B, Havasi C, Lieberman H, Picard R (2012) Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans Interact Intell Syst 2(3) Article 18 MIT Media Lab
14. Nahar V, Unankard S, Li X, Pang C (2012) Sentiment analysis for effective detection of cyberbullying. In: Proceedings of the 14th Asia-Pacific international conference on web technologies and applications LNCS 7235, Springer, Berlin, pp 767–774
15. Nahar V, Al-Maskari S, Li X, Pang C (2014) Semi-supervised learning for cyberbullying detection in social networks. LNCS 8506. Springer, Switzerland, pp 160–17
16. Serra SM, Venter HS (2011) Mobile cyber-bullying: a proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness. Paper presented at Information Security South Africa, Pretoria, South Africa
17. Li M, Tagami A (2014) Study of contact network generation for cyber-bullying detection. Paper presented at 28th international conference on advanced information networking and applications workshops
18. Galán-García P, Puerta JG, Gómez CL, Santos I, Bringas PG (2014) Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. Paper presented at advances in intelligent systems and computing, Springer, Switzerland. vol 239, pp 419–428
19. Fire M, Kagan D, Elyashar A, Elovici Y (2014) Friend and Foe? Fake profile identification in online social networks. J Soc Netw Anal Min 4(1):1–23
20. Fong A, Zhuang Y, He J (2012) Not every friend on social network can be trusted. Paper presented at IEEE international conference on future generation communication technology (FGCT), London
21. Perez C, Lemercier M, Birregah B (2013) A dynamic approach to detecting the suspicious profiles on social platforms. In: Paper presented in IEEE international conference on communications. Budapest
22. Conti M, Poovendran R, Secchiero M (2012) FakeBook: detecting fake profiles in online social networks. Paper presented in IEEE international conference on advances in social networks analysis and mining (ASONAM), Istanbul
23. Liu H, Lim E-P, Lauw H, Le M-T, Sun A, Srivastava J, Kim YA (2008) Predicting trusts among users of online communities—an Epinions case study. In: Proceedings of the 9th ACM conference on electronic commerce. Chicago, pp 310–319
24. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on machine learning, ACM, Pittsburgh, pp 233–240