International Journal on
Advanced Science
Engineering
Information Technology

Vol.10 (2020) No. 3
ISSN: 2088-5334

Detecting the Usage of Vulgar Words in Cyberbully Activities from Twitter

Nursyahirah Tarmizi[a,*], Suhaila Saee[a], Dayang Hanani Abanag Ibrahim[a]

[a] *Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.*
*E-mail: *syahirahmizi93@gmail.com*

Abstract— **Nowadays, nearly all people utilize the device which is connected to Internet. People are accustomed to the use information technology devices in their daily life to interact with other people. Currently, many social media platforms such as Facebook, Twitter, Instagram, and YouTube are becoming popular. This study selected Twitter platforms, which is started to gain popularity. By the rapid growth of users signing up for Twitter accounts, at the same time, cybercrime started to bloom each year in social media platforms. Cyberbully is one of the cybercrime practices which had caused a significant impact on the targeted victims. The victims experienced social pressure, which they need to bear each day while the bullies stayed free behind the veil of anonymity. This study aims to identify the common vulgar words used by the cyberbullies on Twitter. Also, this study is subject to produce essential features of Twitter based on the collected tweets. The evaluation in this study includes the occurrences of the vulgar word perpetrated by the cyberbullies from Twitter. This study detected the usage of vulgar words in cyberbully activities on Twitter platform. A list of vulgar words were extracted and evaluated from a corpus of 50 Twitter users who posted a various number of tweets. The vulgar words detection in the tweets enable the tracking process of the cyberbully activities. In the evaluation section, we discussed how the usage of the vulgar words would define the user's earnestness in doing the cyberbully activities in the Twitter. This study shows there are users with a low number of tweets have a high number of vulgar words occurrences, while other users with high numbers of tweets but less number of vulgar words occurrences. The information collected in this study is expected to assist marking users with a high number of vulgar words occurrences who tend to have high possibilities in doing cyber-bully activities.**

Keywords— **cyberbully; detection; Twitter; social media; vulgar words.**

I. Introduction

Most people nowadays use the Internet and information technology devices in their daily lives to communicate and socialize with others. Presently, social media platforms such as Facebook, Twitter, Instagram, and YouTube are popular socializing tools. People can get connected using the platforms without any distance or time limitations [1]. Social media are popular, especially among youngsters, as they can share or update their daily life activities instantly anytime and anywhere. The online businesses that could be done through social media include updating the status or posting a public message, and photos are also commenting on other timelines [2]. However, with the evolution of social media, the usage of online activities often being used negatively by individuals that find social media platforms as an attractive vehicle to perform illegitimate activities such as cybercrime.

Cybercrime or computer crime is defined as the use of the computer as a tool to further illegal ends, for example, committing fraud, trafficking child pornography, or violating privacy [3]. Twitter as a microblogging platform now has significantly grown to a truly global service that composes of millions of users, which also attracts cyber-criminals to cause cybercrime, including fraud, fake news, and cyberbully [4]. Microblog is a type of blog where users can post a small piece of digital content to share their opinions on the Internet [5]. While cyber-bully is a kind of cybercrime that has grown into a major problem as the usage of online communication and social media keeps increasing [6].

Cyberbullying can be defined as the use of communication technologies to send or post text or images by an individual or group of people that is intended to hurt or embarrass others [7]. Cyberbullying often includes behaviors as opposed to traditional bullying, where the activities involve a direct extension of face-to-face bullying. Although cyberbullying and traditional bullying do share some common ground, cyberbullying constitutes its characteristics [8]. Correspondingly, cyberbully activities are carried out by youth who bully face-to-face and are directed towards the same victims in the same social network circumstances [9].

Moreover, cyber-bully can cause social pressure to targeted victims. As a result, there is a high tendency for victims to experience mental and psychiatric disorders [6].

Later, if the pressure is unbearable, it may lead the victims to attempt or commit suicide. Although there are serious convictions that will be charged towards the cyberbullies, including school discipline, litigation, and criminal prosecution yet, the numbers of cyberbullying cases keep to increase each year [1]. Zainudin et al. stated that there are three reasons why bullies choose to carry out cyberbully activities online:

- The anonymity circumstances that allows the bully to remain anonymous
- Less bravery and courage are needed
- Provides a delusion that the bullies will not get caught and face legal action

Therefore, it is crucial to track down the person who is responsible for perpetrating the cyberbully activities. The primary fundamental factor in tracking the cyberbullies is to study the behavior of the users of social media and monitor their activities online [10]. Thus, understanding user behavior, especially in the context of cybercrime, is to look up the words or vocabulary that are commonly used to humiliate others.

In particular, the study is to understand how the usage of vulgar words enables cyberbully through social media. This paper chooses to focus on analyzing the usage of vulgar words in tweets using Twitter for the following reasons. First, cyberbully activities such as flaming, denigration, and harassment involve high usage of vulgar words or profane words to hurt and terrorize the victim [1]. Second, Twitter is one of the social network sites where cyberbullies use as a cyberbullying tool to attack the victim [11]. Three, Twitter is a highly popular and rapidly growing social network with over 300 million registered users worldwide as of April 2018 [12]. Finally, Twitter provides publicly accessible data.

Meanwhile, the second objective is to produce Twitter's native features that are based on the cyberbullying context. Native features of Twitter, for example, URLs and hashtags, are normalized to standard tag (i.e., “https://www.twitter.com” is changed to “URL” tag). The original tweets are simplified and replaced. The usage of vulgar words in cyberbully activities via social media platforms is discussed. Also, two different techniques for the cyberbully detection system are reviewed.

A. The Usage of Vulgar Words in Cyberbully Activities

The goal of cyberbullies is to harm, dishonor or embarrass a victim through ‘repeated’ acts such as posting inappropriate messages or spreading rumors about the victim online [13]. Henceforth, cyberbullies hold power to embarrass or terrorize a victim before an entire community online [14]. These cyber bullies can hide their true identity behind the veil of anonymity in social media and remain undiscovered.

The element of perceived anonymity via online activities facilitate the cyberbullies to appear safe and secure. Anonymity is defined as the state of being unknown to most people. The issue of anonymity has made the situation tougher, especially to law enforcement in gathering sufficient evidence for jurisdiction [15]. Regardless of the anonymity nuisance, there are attempts and proposed solutions to surpass the cyberbully activities in social media. There are six main types of cyberbully activities happen in online social networks [1].

- Flaming: An activity of harsh argument that usually takes place in an instant message, email, or chat rooms by using vulgar words in provoking or offensive messages to someone by a group of people.
- Trolling: Posting provocative messages to create, upset and baiting people to fight.
- Denigration: Unfairly criticizing a person by posting cruel gossips or rumors to damage the victim reputation.
- Harassment: The ability to interact or contact other people either with or without permission. Include speech abuse, self-harm, being rude, and post sexual content to an online user.
- Masquerade: The bully creates a fake profile ID and pretending to be someone else and keep the bullying occur. The culprits are anonymous, yet they can still harass the victims.
- Cyberstalking: Stalking other people information to make a false accusation, monitoring, identity theft, threats, and create data destruction or manipulation.

B. Cyberbully Detection System using Machine Learning Techniques

Raisi and Huang proposed a cyberbully detection model by using the user-vocabulary consistency based on the curse and bully words. The bully words which contain in the seed dictionary (bullying indicator) act as the indicator to measure the score of the ‘bully’ and the score of the ‘victim’. The bully score is used to measure how much a user tends to bully others. On the other hand, the victim score is used to measure how much a user tends to be bullied by others. They use Twitter and Ask.fm as the dataset to evaluate their work.

Meanwhile, a study has developed an automated tool to detect cyberbully activities in a forum post [16]. The aggressiveness and the anonymity of the posts are labeled manually. Later, the labels are used to identify the aggressiveness of the attacks, which includes both the attacker and the defender. This paper uses the second-person pronouns and profanity words as the aggressiveness indicators. They deploy the text-matching techniques which utilize the profanity words, sentence structure, and pronouns as the features. The list of profanity words is obtained from an open-source dictionary. The automated tool is developed to help in identifying the aggressive attacks from the forum posts.

A study has proposed a solution for detecting textual cyberbullying by building individual topic-sensitive classifiers [17]. They deploy Lexical Syntactic Features (LFS) to detect insulting content and identify the potential offensive users in social media. Besides that, they also distinguish the contribution of pejoratives/profanities and obscenities in determining offensive content. The dataset consists of YouTube comments that are related to sensitive issues for an instance of race and culture, sexuality, and physical appearance. Apart from that, the comments have also been extracted based on a list of profane words that indicated the harsh and rude comments. The features include the user writing style, structure, and unique content related to cyberbullying to predict the user's potential to post such insulting content.

C. Cyberbully Detection System using Text Matching Techniques

Research studies the properties of the cyber-bullies and aggressors by finding out the features that can be used to distinguish the bullies and aggressors from regular users [18]. This paper state that to label the aggressiveness and bullying behavior in Twitter, the labels are obtained from the cloud sourcing platform, which utilizes the human annotations in labeling the classes. According to this study, it shows that the offensive users and the bullies tend to attack in a short burst, specific users, or targeted groups. They have extracted three feature sets from a collection of tweets, which consist of the text, network-based, and user-based attributes. Among the three features, network-based features comprise of the user connectivity are proven as the most compelling features in detecting the aggressiveness of the user behavior. The corpus used to consist of 1.4 million tweets that have been collected for three months.

All the above-mentioned studies covered the background of cyberbullying and the cyberbully detection systems that have been proposed to thwart the cyberbully activities in social media and forum post. A study had utilized the textual features to detect the aggressiveness of the post or tweets [18]. Besides, this paper proved that the other features that are native to the social network platform are more effective in detecting the aggressiveness behavior of the user. Some studies deployed a different approach where they use machine learning technique [7],[17], [19] and another study had deployed a text-matching technique to enhance the detection system for cyberbullying and aggression [18].

II. Material and Method

A. Methodology

In this paper, a workflow to detect the usage of vulgar words in cyberbully activities from Twitter is proposed. Fig. 1 depicts the workflow to detect the usage of vulgar words in cyberbully activities from Twitter.

The detection involves analyzing collected tweets that are extracted from Twitter which contain vulgar words in the tweets. According to Fig. 1, there are three main processes that are carried out throughout the workflow which are:

1) Data Collection: To collect the tweets that contain vulgar words from Twitter, a data crawler is used in this experiment. A dictionary of vulgar words is obtained from a website namely noswearing.com. The dictionary contains 370 swear and curse words. Then, the dictionary is inserted into the Twitter crawler namely, Tweepy is a python library used to crawl the tweets using the Twitter API. Basically, Tweepy provides a method that tracks down the keywords in the dictionary and produces a list of all current tweets that contain the keywords in them from the Twitter.

2) Pre-processing: After the tweets are retrieved from Twitter, more tweets are extracted from the users' timelines. During pre-processing, there are specific data that are removed, such as:

- Retweet messages – messages that are retransmitted by another user. This message does not contain the user information.
- 'RT@' messages – retweeted messages with user mention
- Non-English tweets – tweets that are not in English are removed since our focus is on English-language tweets only.

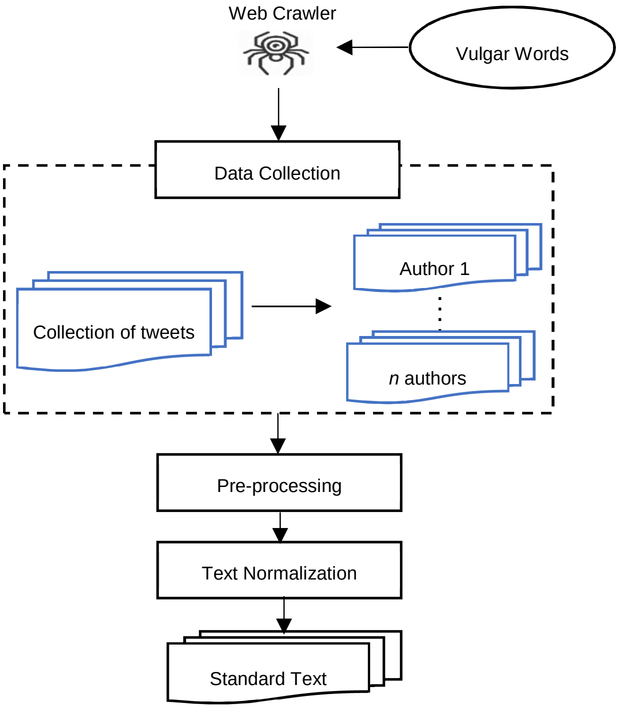


Fig. 1 The workflow of detecting the usage of vulgar words in cyberbully activities from Twitter

3) Text Normalization: After undergoing pre-processing, the collection of tweets are normalized. The normalization process includes replacing some sparse characteristics such as numbers, date and time, URL links, user references, and hashtags. This process is needed to simplify the representation of text and to reduce data sparsity [20]. The process is done by appending the original text and replace it with standard tags that represent the original content. The following example below shows tweets that undergo a normalization process.

- Before normalization:
  Tweet 1: "You can count on me like 123"
  Tweet 2: "Can't wait this coming 9/5/2018"
- After normalization:
  Tweet 1: "You can count on me like NUM"
  Tweet 2: "Can't wait this coming DAT"

B. Experimental Setup

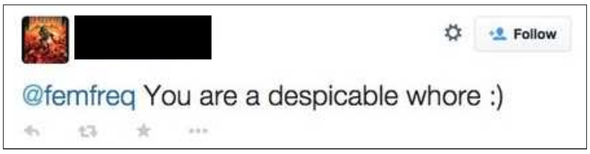
This paper aims to detect cyberbully activities on Twitter by tracking down tweets that might contain insult and swear words. The experiment is conducted in a controlled environment where all the data are analyzed on one single machine. This is to ensure the consistency in the results of the performed test. The configuration of the machine is as follows:

TABLE I
THE INFORMATION OF THE COMPONENTS

ID	Component	Info
01	CPU	Intel Core2 Quad
02	RAM	4.00 GB
03	OS	32-bit Windows OS
04	Python	Version 2.7.15
05	Tweepy	Version 2.0.7

1) Tweet Taxonomy: Twitter is a microblogging social media platform where users post using short messages known as “tweets”. Tweets consist of short messages with 280 characters or less, including text, images, videos, locations and emojis (small digital images or icons to express ideas). Moreover, there are native elements in Twitter such as links, hashtags and user references that can be included in the tweets [20]. As for the shared links, the links are counted as 23 characters, no matter how long the link is. Hashtags, on the other hand, are used to categorize the tweets by keyword. The users use the hashtag symbol ‘#’ before the keyword or phrase for example, ‘#worldcup’. The hashtag feature in Twitter helps to categorize the tweets and list related tweets that have the same keyword or phrase. A single tweet may consist of a combination of those mentioned above elements.

2) Input Dataset: The data which are the tweets that are obtained through crawling might contain curse words and vulgar words in them. The following example in Fig. 2 show a tweet that contain vulgar words posted by a user to the victim.



[redacted]

Follow

@femfreq You are a despicable whore :)

Fig. 2 Example of tweet containing vulgar words

The usage of the vulgar words in tweets makes these tweets especially prone to cyberbullying on Twitter. By observing and tracking the usage of vulgar by the users, this can assist in studying cyberbully activities in Twitter. The tweets with vulgar words are deemed to include as offensive messages that will lead to cyberbullying. Therefore, this can help in the development of communicative model for cyberbullying detection. There are offensive or rude words will make a tweet more likely to be labelled as an offensive message which leads to cyberbullying. To leverage the information gathered, we identified a list of insult and curse words, posted on a website, www.noswearing.com. A list of 357 vulgar terms was downloaded, and a dictionary of these words was created and inserted into the crawler to crawl the tweets.

In this experiment, 50 tweets that contain vulgar words in them are randomly extracted using a twitter crawler. The crawler is used to download random tweets in real-time. After randomly extract 50 tweets from the Twitter, a list of users that posted the tweets is obtained through their ‘user_id’ respectively. The ‘user_id’ parameter holds a unique identifier for that particular user and the id is used to stream the most recent 3200 tweets posted by that user. The collection of remaining tweets posted by 50 users are extracted from their timeline, respectively, and are stored as the corpus. The corpus consists variety number of tweets posted by 50 different users. Table 2 sums up the statistic about the corpus. It is important to notice that the corpus consist of concise documents with a maximum length of 280 characters and only correspond to English tweets.

TABLE II
STATISTICAL DESCRIPTION OF THE CORPUS

Description	Amount
Total user	50
Total number of tweets	69,624
Total number of words (tokens)	988,270
Total vulgar words occurrences	29,701
Average number of tweets	1,392
Average number of words	988,270
Average number of vulgar word occurrences	594
Maximum number of vulgar words occurrences	3187
Minimum number of vulgar words occurrences	9

III. RESULT AND DISCUSSION

This section reports the results of the experiment and the analysis of the findings.

A. *Corpus Analysis*

Figure 3 below summarizes the statistical description of the corpus. The graph in Fig. 3 contains the information about the number of users (x-axis). In the y-axis, there are two kinds of information where the primary y-axis represents the number of words while the secondary y-axis depicts the vulgar words occurrences information.

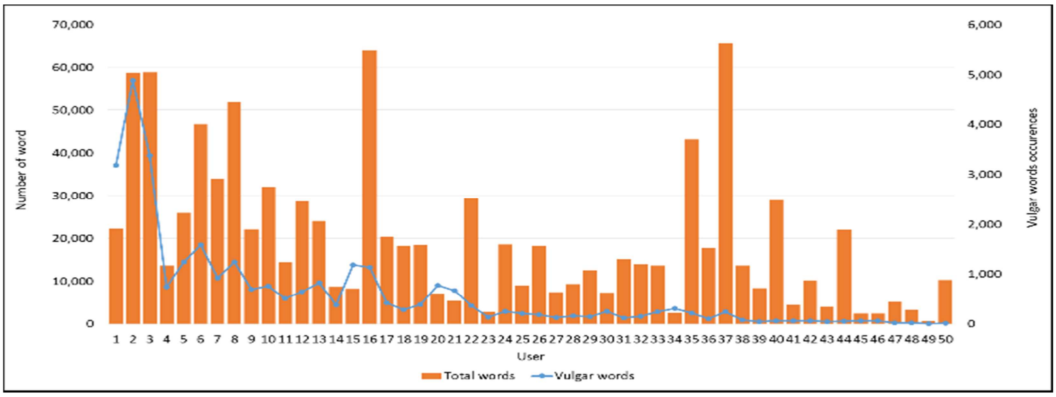


Fig. 3 Users information graph

Based on Fig. 3 above, the graph depicts that the first user has the highest value of the vulgar words in the total of 22,316 words with 3,187 vulgar words occurrences. The average vulgar words occurrences by user 1 is 455 words. The last bar on the graph indicates the information of last user, which is user 50. User 50 has the highest total number of words (10,314 words) in between user 45 to user 49. Yet, the occurrences of the vulgar word for user 50 is only 9 vulgar words and the average occurrences is only 0.4.

Like the user 50, user 37 also has the same situation where user 37 has the highest number of words with 65,671 words. However, the number of vulgar words occurrences is only 244 words, which shows the average of vulgar words occurrences is only 9 vulgar words. For user 3 and 4, both users have total words that are larger than user 1 with 58,789 words and 58,931 words, respectively. However, the average occurrences by both users are lesser than user 1, with 264 average vulgar words occurrences for user 2 and 182.3 average vulgar words for user 3.

The analysis of the graph suggests that, if the total number of words is big, this does not mean that the user have high number of vulgar words occurrences. There are users with a total number of words below than 10,000 but have a high number of vulgar words occurrences. Therefore, the only way to determine which user has the most usage of vulgar words is by calculating the ratio of total number of words and vulgar words occurrences of that user. As a result, users with a high number of average vulgar words occurrences are the possible perpetrators of cyberbully activities.

B. Vulgar Words Evaluation

During the classification process, there are words that are misclassified as vulgar words. The graph in Figure 4 below depicts the number of correct and incorrect classification of the vulgar words. The process of classification is done semi-automated. The vulgar words are extracted based on the vulgar words dictionary as mentioned in section III as well as the process of extraction is mentioned in section IV. After the vulgar words' extraction process is completed automatically, the classification process is done manually.

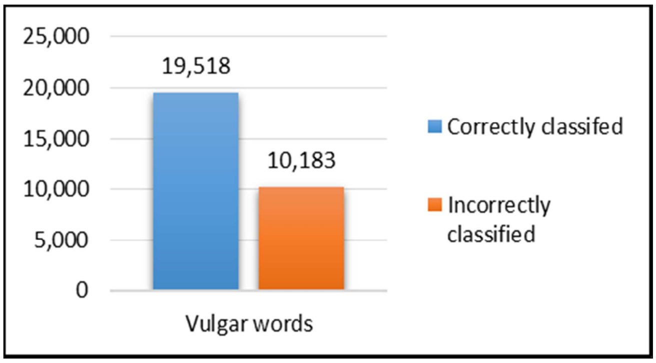


Fig. 4 Graph of correctness in classifying the vulgar words

During the classification process, we have found out that there are words that do not belong to the vulgar words class but extracted as the vulgar words. The graph in Fig 4 depicts the number of correct and incorrect classification of the vulgar words. From a total of 29,701 vulgar words occurrences, there are 19,518 words that are correctly classified as vulgar word. However, the remaining 10,183 words are misclassified as vulgar words. The evaluation of the usage of vulgar words for each user is done by computing the precision and error rate value based on the confusion matrix value (TP = true positive and FP = false positive). Table 3 shows the precision value. The calculation involves the fraction of the predicted positive is positive (TP) is counted.

TABLE III
THE EVALUATION RESULTS OF VULGAR WORDS OCCURRENCES

Average TP rate	Average FP rate	Average Precision	Average Error Rate
0.6571	0.3429	0.6976	0.3024

The result of precision represents the exactness of classifying the vulgar words. While, the error rate calculation is to measure the misclassification of vulgar words. Whereas, TP-value (truth positive) indicates the correct classification of vulgar words. FP-value (false positive) indicates words that are misclassified as the vulgar words. The results in Table 3 indicates that the average rate of precision is 70%. Previous study obtained a good result of precision (91%) in detecting cyberbully activities using words matching techniques [6] and obtained 90% accuracy using Naïve Bayes classifier to classify the cyberbully activities in Twitter [7]. However, in this paper, the evaluation covers only the data collection phase. Therefore, the result of the classification is lower.

IV. CONCLUSION

To conclude, the aim of this paper is to detect the usage of vulgar words in cyberbully activities from Twitter. A list of vulgar words has been extracted and evaluated from a corpus of 50 Twitter users that posted a various number of tweets. The detection of vulgar words in the tweets facilitate the process of tracking the cyberbully activities from this platform. In the evaluation section, we have discussed how the usage of the vulgar words will define the seriousness of the users in conducting the cyberbully activities in the Twitter.

Our analysis showed that there are users with a low number of tweets have a high number of vulgar words occurrences. While, there are users with high numbers of tweets but less number of vulgar words occurrences. The information gathered in this experiment can assist to mark users with a high number of vulgar words occurrences. These users with a high number of vulgar words occurrences tend to have high possibilities in conducting cyber-bully activities.

Although the precision rate is only average with almost 70%, yet the detection of vulgar words in tweets is still high with more than 29 000 vulgar words occurrences from 69 000 total number of words. This shows that the Twitter users are actively using vulgar words in their tweets to convey messages. Besides, our experiment indicates that our method can detect new vulgar words vocabulary. Thus, the information gathered from this experiment will be used later in the next process of identifying the author of an anonymous tweet containing cyberbullying sentiment in it.

The evaluation of this experiment can be used as an aid in improving the accuracy of the identification system later for social media forensics.

References

[1] N. M. Zainudin, K. H. Zainal, N. A. Hasbullah, N. A. Wahab and S. Ramli, "A Review on Cyberbullying in Malaysia from Digital Forensic Perspective," in 2016 International Conference on Information and Communication Technology (ICICTM), Kuala Lumpur, 2016.

[2] F. Martin, C. Wang, T. Petty, W. Wang and P. Wilkin, "Middle School Students' Social Media Use," Educational Technology & Society, vol. 21, no. 1, p. 213–224, 2018.

[3] R. Sabillon, J. Cano, V. Cavaller and J. Serra, "Cybercrime and Cybercriminals: A Comprehensive Study," International Journal of Computer Networks and Communications Security, vol. 4, no. 6, p. 165–176, 2016.

[4] A. Alexandrou, "Cybercrime," International and Transnational Crime and Justice, vol. 10, p. 61, 2019

[5] I. L. Liu, C. M. Cheung and M. K. Lee, "User satisfaction with microblogging: Information dissemination versus social networking," Journal of the Association for Information Science and Technology, vol. 67, no. 1, pp. 56-70, 2016.

[6] S. A. Özel, E. Saraç and S. Akdemir, "Detection of cyberbullying on social media messages in Turkish," in Computer Science and Engineering (UBMK), 2017 International Conference, Antalya, 2017.

[7] M. A. Al-garadi, K. D. Varathan and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," Computers in Human Behavior, vol. 63, pp. 433-443, 2016.

[8] P. Sara and V. Heidi, "An investigation of short-term longitudinal associations between social anxiety and victimization and perpetration of traditional bullying and cyberbullying," Journal of youth and adolescence, vol. 45, no. 2, pp. 328-339, 2016.

[9] R. Festl and T. Quandt, "The role of online communication in long-term cyberbullying involvement among girls and boys," Journal of youth and adolescence, vol. 45, no. 9, pp. 1931-1945, 2016.

[10] A. M. Chandrashekhar, M. G. S and A. D. K, "Cyberstalking and Cyberbullying: Effects and prevention measures," Imperial Journal of Interdisciplinary Research (IJIR), vol. 2, no. 3, pp. 95-102, 2016.

[11] Ditch the Label, "The Anual Bullying Survey 2017," Ditch the Label, United Kingdom, 2017.

[12] "The Statistic Portal," Statista, April 2018. [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. [Accessed 7 July 2018].

[13] G.Sarna and M.Bhatia, "Content based approach to find the credibility of user in social networks: an application of cyberbullying," International Journal Of Machine Learning and Cybernetics, vol. 8, no. 2, pp. 677-689, 2017.

[14] R. Garett, L. R. Lord and S. D. Young, "Associations between social media and cyberbullying: a review of the literature," Mhealth, vol. 2, no. 8, p. 46, 2016.

[15] A. Usha and S. M. Thampi, "Usha, A., & Thampi, S. M. (2017, December). Authorship Analysis of Social Media Contents Using Tone and Personality Features," in International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, Cham, 2017.

[16] B. Mascotto, "Exploring the Impact of Anonymity on Cyberbullying in Adolescents:" University of Victoria, Victoria, 2008.

[17] E. Raisi and B. Huang, "Cyberbullying Identification Using Participant-Vocabulary Consistency," in ICML Workshop on #Data4Good: Machine Learning in, New York, 2016.

[18] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter," in Proceedings of the 26th International Conference on World Wide Web Companion, Perth, 2017.

[19] P. Gal´an-Garc´ıa, J. G. d. l. Puerta, C. L. G´omez, I. Santos and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," Logic Journal of the IGPL, vol. 24, no. 1, pp. 42-53, 2016.

[20] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. Carvalho and E. Stamatatos, "Authorship Attribution for Social Media Forensics," IEEE Transactions on Information Forensics and Security, pp. 5-33, 2017.