JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(1): 24-34, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.1.2

Identification of Profane Words in Cyberbullying Incidents within Social Networks

Wan Noor Hamiza Wan Ali
Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: p93244@siswa.ukm.edu.my

Masnizah Mohd*
Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: masnizah.mohd@ukm.edu.my

Fariza Fauzi
Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: fariza.fauzi@ukm.edu.my

ABSTRACT

The popularity of social networking sites (SNS) has facilitated communication between users. The usage of SNS helps users in their daily life in various ways such as sharing of opinions, keeping in touch with old friends, making new friends, and getting information. However, some users misuse SNS to belittle or hurt others using profanities, which is typical in cyberbullying incidents. Thus, in this study, we aim to identify profane words from the ASKfm corpus to analyze the profane word distribution across four different roles involved in cyberbullying based on lexicon dictionary. These four roles are: harasser, victim, bystander that assists the bully, and bystander that defends the victim. Evaluation in this study focused on occurrences of the profane word for each role from the corpus. The top 10 common words used in the corpus are also identified and represented in a graph. Results from the analysis show that these four roles used profane words in their conversation with different weightage and distribution, even though the profane words used are mostly similar. The harasser is the first ranked that used profane words in the conversation compared to other roles. The results can be further explored and considered as a potential feature in a cyberbullying detection model using a machine learning approach. Results in this work will contribute to formulate the suitable representation. It is also useful in modeling a cyberbullying detection model based on the identification of profane word distribution across different cyberbullying roles in social networks for future works.

Keywords: cyberbullying, profane words, cybercrime, harassment, social network, machine learning

Received: October 19, 2020 **Accepted:** February 21, 2021

***Corresponding Author:** Masnizah Mohd
https://orcid.org/0000-0001-8908-8755
E-mail: masnizah.mohd@ukm.edu.my

1. INTRODUCTION

The rapid development of information and communication technologies and the great proliferation of social networking sites (SNS) have made social media a popular socialization tool, especially among teenagers. Since the introduction of these SNS such as Facebook, Instagram, Twitter, and YouTube, they have been able to attract the attention of millions of users to the point where these SNS have become a necessity in life (Tarmizi et al., 2020). According to Boyd and Ellison (2007), social networks are defined as web-based services that allow individuals to build public or semi-public profiles in a limited system, and to clearly know the list of other users with whom they connect as well as to view their contact lists. Thus, the use of the Internet as a medium of communication makes SNS an important socialization tools to share information, exchange opinions, interact with each other, and to conduct timeless and borderless online business (Zainudin et al., 2016). However, SNS could give negative implications if people use SNS for cybercrime purposes (Sheeba et al., 2019).

1.1. Cyberbullying Detection

One of the cybercrimes that usually occur on SNS is cyberbullying (Nandhini & Sheeba, 2015). Cyberbullying has been defined in multiple ways but have the same characterization where it takes into account the people involved, the kind of medium, and how cyberbullying occurs. Hinduja and Patchin (2018) defined cyberbullying as "willful and repeated harm inflicted through the medium of electronic text." Based on this definition, they did not mention cyberbullying that happened based on the specific age group of users, in contrast to other researchers who believe that cyberbullying happens only among students such as children and teenagers (Sutton, 2011). When it comes to adults, cyberbullying can be defined as cyber harassment or cyberstalking, besides cyber incivility (Giumetti et al., 2012; Sutton, 2011). However, cyberbullying is an aggressive behavior by an individual or group of people known as harassers towards another individual (i.e., victim) using the Internet as a medium. Harassers are able to use various ways to bully the victim such as spreading rumors and spreading information related to the victim for the purpose of embarrassing him/her. The people involved in cyberbullying incidents can be classified as a harasser, victim, assistant, defender, reinforcer, and accuser (Salmivalli, 1999; Salmivalli et al., 1996; Xu et al., 2012). Table 1 shows the description of these six roles. These individuals may play different roles in a single cyberbullying incident (Xu et al., 2012). Van Hee et al. (2018) only consider four roles: 1) harasser, 2) victim, 3) assistant, and 4) defender. They refer to the assistant and defender in cyberbullying incidents as bystanders (specifically, bystander-assistant and bystander-defender, respectively).

The impact and frequency of cyberbullying incidents can intensify or lessen depending on the roles involved (Moxey & Bussey, 2020). The existence of the roles beside harasser and victim has been recognized as important (Bastiaensens et al., 2014; Salmivalli, 2010). If a cyberbullying incident involves only the harasser and victim, the issue is likely to be cut short. However, if the incident involves roles such as assistant and reinforcer, this situation can worsen and become more dangerous, as the harasser has received support from other online social network users, and does not consider the cyberbullying wrong, and will continue to threaten the victim. Nonetheless, if there are users playing the role of defender involved, it is likely that the incident will not be extended because the harasser has no support from the other online social network users (Salmivalli, 2010). Thus, individual involvement in cyberbullying incidents is noteworthy in cyberbullying detection.

Discriminative features are required to develop a cyberbullying detection model by using a supervised machine learning approach. In general, basic steps in modeling the

Table 1. Description of roles involved in cyberbullying

Role	Description
Harasser	Person who did the bullying across online social networks
Victim	Person who was the target of harasser
Assistant	Helps harasser to do the bullying
Defender	Defends victim from harasser
Reinforcer	Indirectly involved in cyberbullying but encourages harasser and gives impetus for continuation, such as cheering or laughing
Accuser	Accusing someone as the harasser

cyberbullying detection are started from data collection, data annotation, data pre-processing, feature extraction, and classification (Talpur & O'Sullivan, 2020). However, collecting data from online social networks is not an easy task. According to Salawu et al. (2017), the challenge lies in the form of privacy and ethical matters which most researchers have to face before the data can be harvested, especially when it involves data related to user information such as age and gender. Besides this, different online social networks provide different metadata, and researchers should determine the detection task before selecting the data source to be used. However, the corpus should be annotated after collecting the same from social networks. The data annotation is the process where raw data will be identified into meaningful and informative annotation to classify text. Meanwhile, data pre-processing is one of the crucial steps in cyberbullying detection because we need to remove outliers and standardize the data before it can be used to build the cyberbullying model during classification process. A few examples of data pre-processing are tokenization, stemming, lemmatization, remove numbering, and stopwords removal.

Feature extraction in cyberbullying detection is dependent on the corpus data that have been harvested. As an example, the corpus may or may not contain metadata of elapsed time between comments, followers, following, time of posting, and other data. These metadata may be required to identify the interaction between users of the online social network site. Salawu et al. (2017) categorized four types of features used in cyberbullying detection: 1) content-based (e.g., cyberbullying keywords, profanity, pronouns, and n-grams), 2) sentiment-based (e.g., emoticons), 3) network-based (e.g., number of following, number of followers), and 4) user-based (e.g., age, gender). Content-based features are commonly used features by researchers in cyberbullying detection. In addition, the profanity feature is the highest feature that has been used across cyberbullying detection studies as compared to other features in the content-based category, since cyberbullying is associated to an aggressive behavior (Salawu et al., 2017). Finally, we use an algorithm as classifier in the classification process as the final step for cyberbullying detection. Common algorithms usually implemented in supervised machine learning are the Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, and Linear Regression (Muneer & Fati Mohamed, 2020).

However, state-of-the-art of cyberbullying detection using a machine learning approach has focused on profane words as a feature rather than utilizing the profane features according to four roles (harasser, victim, assistant, and defender) involved in cyberbullying incidents. This has motivated us to identify profane words from the corpus to analyze the word distribution based on four roles involved in cyberbullying incidents within social network. However, a limitation of this work is that we only identify lexical profane words in accordance to a profane word dictionary. We addressed the following research questions:

1. Which role used profane words with the highest frequency based on word distribution?
2. What are the commonly used words by users in cyberbullying incidents?

We hypothesize that identifying word distribution for each role involved in cyberbullying detection can give useful insights for us to better extract profanity features from the corpus for use in the classification process. Results from the experiment will guide us for future work in cyberbullying detection using a machine learning approach. The contribution of this work is noteworthy, as it focuses on profane words distribution based on four different roles involved in cyberbullying detection. The results from this work will be implemented in formulating the representation of features in order to develop a model of cyberbullying detection. We also combined two different external sources of profane word dictionaries with the purpose to strengthen the results of the experiment.

The remainder of this paper is structured as follows: Section 2 presents the related work on automatic cyberbullying detection using profanity as a feature. The methodology in Section 3 describes the workflow for the identification of profane words. Next, we present the results from the experiment in Section 4. Meanwhile, Section 5 discusses the analysis of profane word distribution. Finally, Section 6's conclusion and future work concludes this paper.

2. RELATED WORK

Nowadays, profane words are commonly used in daily face to face conversations as well as on the Internet, especially in online social networks (Laboreiro & Oliveira, 2014). A profane word is a curse word that is used with offensive meaning and usually used in cyberbullying incidents and hate speech (Laboreiro & Oliveira, 2014). Profanity is a common feature used in cyberbullying detection. State-of-the-art works on cyberbullying detection that used profanity feature to detect cyberbullying in on-

line social networks include research works by Dinakar et al. (2011), Van Hee et al. (2018), Kontostathis et al. (2013), and Al-garadi et al. (2016).

Dinakar et al. (2011) have focused on cyberbullying detection based on textual communication that involved three main sensitive topics, which are sexual, race/culture, and intelligence. By using a corpus from YouTube, they annotated the corpus that consists of negative and profane words according to those three topics. However, they stated that if a comment in a YouTube video contains negative or profane words, it does not necessarily mean that the comment can be labeled as a cyberbullying incident. For example, "I'm disgusted by what you said today, and I never want to see you again" is difficult to classify as cyberbullying even though that sentence indicates the lexicon 'disgusted.'

Profane words are also identified by Al-garadi et al. (2016) in cyberbullying detection because they believe it is a sign of offensive behavior. They extracted profane words based on a dictionary of profanities which is compiled in previous research works by Reynolds et al. (2011) and Wang et al. (2014). In addition, a study by Chatzakou et al. (2017) uses an external resource, the 'hatebase' database (https://hatebase.org) to determine the 'hateful' score for the collected tweets on a scale [0,100], where 0 indicates no hatred and 100 indicates hatred. However, they stated that profane words are not suitable for classifying tweets as hateful or aggressive cyberbullying, as the tweets are short and commonly comprise altered words, emoticons, and URLs. So, they concluded that users who like to harass other users have a tendency to use profane words, but they are not much different from normal user behavior.

The Google profanity list (https://code.google.com/archive/p/badwordslist/downloads) is used as a lexicon dictionary for feature extraction by Van Hee et al. (2018). However, Van Hee et al. (2018) stated that profanity has a low percentage in F1 measure but it still can be a useful feature if it is integrated with other features such as sentiment, n-grams (character), and topic model.

Tarmizi et al. (2020) used the corpus from Twitter to detect cyberbullying activities. The profane words extracted from the 50 Twitter users' corpus is used to track the activities by the users. The extraction is based on an external source, www.noswearing.com. In the classification process, about 10,183 from 29,701 profane words occurrences are misclassified. It may be due to the low number of tweet activities having a high number of profane words and the high number of tweet activities having a low number of profane words.

Current findings showed that profane words were highly used as a feature to detect the cyberbullying that happened in social networks. Nonetheless, there is no research that considers frequency of profane words based on roles involved in cyberbullying incidents. This is important to ensure that feature use in classification processes can build a robust model for cyberbullying detection.

3. METHODOLOGY

We set up an experiment to test the aforementioned hypothesis and answer the research questions. Thus, in this section a workflow to analyze profane word distribution across the different cyberbullying roles in the corpus is proposed as illustrated in Fig. 1. The workflow involves four steps: 1) dataset, 2) data preparation, 3) data preprocessing, and 4) profane word extraction.

3.1. Dataset

The corpus used in this work is the AMiCA Bullying Cyber Dataset, a secondary dataset obtained from Van Hee et al. (2018). The corpus is collected from ASKfm's social network site from April to October 2013. It consists of English and Dutch language. However, only the English language is used in this study. ASKfm is one of the most popular online social networking platforms for teens and adults where users can interact without having to reveal their own identities (Ashktorab et al., 2017). The corpus obtained is in the form of annotations annotated using Brat Rapid Tool Annotation (BRAT) (Stenetorp et al., 2012).

In the annotation process as described in Van Hee et al. (2018), messages collected are presented to the annotators in chronological order within the context of the original content of the conversation. One conversation consists of a series of messages from two or more users. In each conversation, each message is preceded by a ¶ token. The token is intended to facilitate and speed up the annotation process. Firstly, the nature of each message is taken into account whether it is harmful or otherwise. The level of harmfulness is indicative of cyberbullying. Hence, for each message, the annotators determine whether the message contains indication of cyberbullying or not based on three types of scales, namely:

1. Harmfulness Score=0: Messages that do not contain cyberbullying or indication of cyberbullying.
2. Harmfulness Score=1: Message containing cyberbullying or an indication of cyberbullying.

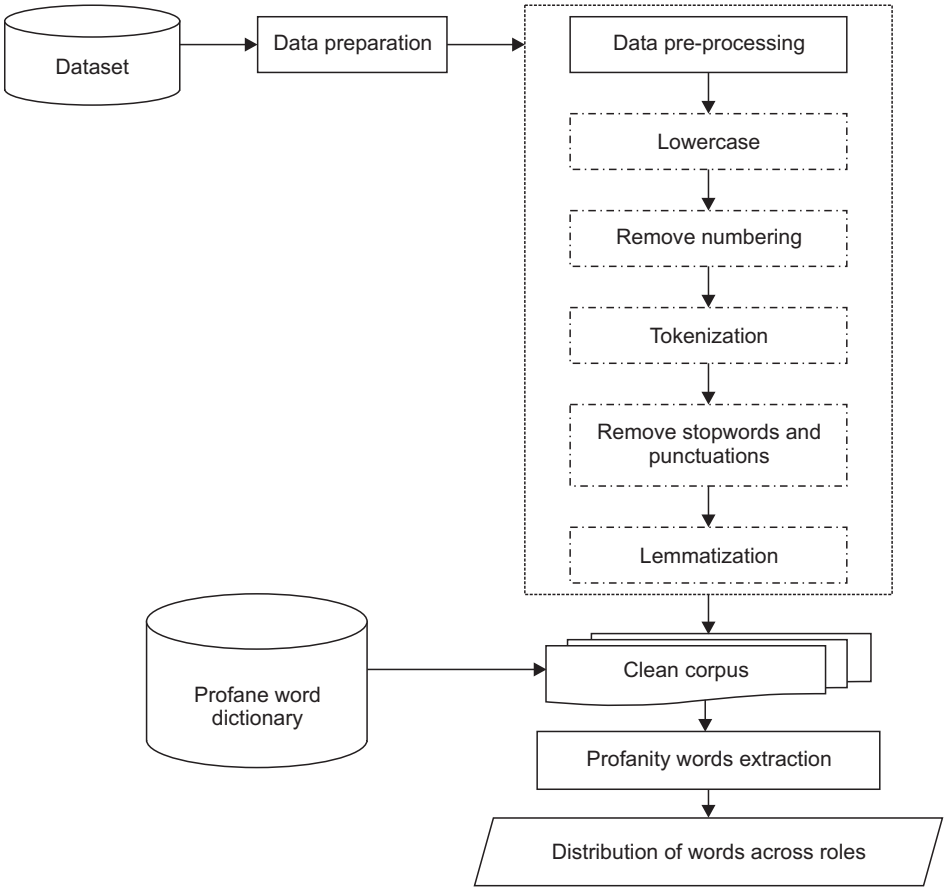


Fig. 1. Workflow of profane words extraction.

3. Harmfulness Score=2: Messages that contain a serious indication of cyberbullying (e.g., death threats).

Secondly, if the message is rated with a harmfulness score of 1 or 2, the annotators must specify the user's role for the messages in the following format: *score_role* i.e., *1_harasser, 2_harasser, 1_victim, 2_victim, 1_bystander_assistant, 2_bystander_assistant, 1_bystander_defender,* or *2_bystander_defender.*

In this work, profane words are extracted from four different roles without any consideration for the harmfulness score. For example, both *1_harasser* and *2_harasser* are simply considered as *harasser* for the extraction of profane words.

3.2. Data Preparation

Data preparation processing is performed upon obtaining the dataset. All files containing the original messages and annotations are unified in one folder. The BRAT reader program is imported and implemented in the Anaconda Spyder platform using Python 3.7 programming language, and the original Python codes from https://github.com/clips/bratreader/tree/master/bratreader are adapted to parse the original messages and annotations. BRAT reader is a Python program for reading and understanding BRAT repositories and for storing files in XML format for easy access (Computational Linguistics Research Group, 2019). The program was developed by Computational Linguistics & Psycholinguistics from the University of Antwerp, Belgium. The original messages and annotation files are loaded into .xml files before being stored and converted to .csv format using the BRAT reader program. Then we imported the converted files into the Anaconda Spyder platform for data pre-processing.

The corpus is then processed for better understanding. Table 2 shows the descriptive statistics of the corpus. The corpus has a total of 120,512 posts. They were classified into two types of post which are cyberbullying posts (5,382) and non-cyberbullying posts (115,130). For the cyberbullying posts, the number of posts is calculated for each of the four roles defined in a cyberbullying incident. As can be observed from Table 2, the corpus is imbalanced (only 4.47% of the corpus are labelled as cyberbullying posts) and this situation will affect the performance

 https://doi.org/10.1633/JISTaP.2021.9.1.2

of standard classifiers and predictive models. This issue will be addressed in our future work.

3.3. Data Pre-Processing

Based on the corpus provided, each original message goes through numbers removal and lowercase conversion. Then, all punctuation and stopwords are removed from the corpus. The next step is the tokenization of the corpus by converting each sentence in the message into a set of words known as tokens for easier comprehension (Sekharan et al., 2018). Finally, lemmatization is performed to obtain the root word for each token. Table 3 shows the pre-processing steps of the original data and processed output using the following example sentence: "NIGGA you look like you are a 60 years old hag, even SLUTS like you!!!"

3.4. Profane Word Extraction

We utilized two different external sources to identify profane words from the corpus according to the four different roles involved in cyberbullying, as follows:

1. Curse word lexicon from https://www.cs.cmu.edu/~biglou/resources/bad-words.txt, that consists of 1,383 lexicons (Zois et al., 2018)
2. Google profanity list https://code.google.com/archive/p/badwordslist/downloads, which consists of 458 lexicons (Van Hee et al., 2018)

Combining both external sources gives 1,841 lexicons and we compiled this into one file before identifying the profane words. The identification of profane words from the corpus is conducted in a controlled environment on a single machine. The reason is to make sure of consistency of performance in order to produce results. We performed countvectorizer for feature extraction, which is used to convert a corpus to a matrix of tokens by making use of Scikit-learn (Pedregosa et al., 2011). Then, we implemented fit_transform of lexicon dictionary, and word distribution for each role is calculated.

4. RESULTS

In this section, we present the findings from an experiment of profane word identification as a part of feature extraction, which could be used in the classification process of cyberbullying detection. Table 4 shows the percentage of profane words distribution across four roles in cyber-

Table 2. Descriptive statistics of ASKfm corpus

Post	Corpus size	Percentage of corpus size (%)	Score_Role	Number of post	Total post each role
Non-cyberbullying post	115,130	95.53	-	-	-
Cyberbullying post	5,382	4.47	1_Harasser	2,886	2,886+693=3,579
			2_Harasser	693	
			1_Victim	1,276	1,276+79=1,355
			2_Victim	79	
			1_Bystander_Defender	391	391+33=424
			2_bystander_Defender	33	
			1_Bystander_Assistant	17	17+7=24
			2_Bystander_Assistant	7	

Table 3. Output for each pre-processing step

Pre-processing steps	Output for pre-processing
Step 1: Remove numbering	NIGGA you look like you are a 60 years old hag, even SLUTS like you!!!
Step 2: Lowercase	nigga you look like you are a years old hag, even sluts like you!!!
Step 3: Remove punctuations and stopwords	nigga hag sluts
Step 4: Tokenization	['nigga', 'hag', 'sluts']
Step 5: Lemmatization	['nigga', 'hag', 'slut']

bullying incidents. A formula to calculate the percentage of profane word distribution is built. The number of profane words indicated the total number of profane words that have been extracted from tokens for each role. The formula as follows:

$$\text{Percentage Profanity Words Distribution} = \frac{\text{Number of Profane Words}}{\text{Total Number of Tokens}} \times 100\%$$

The Harasser role consists of 12,958 tokens from 3,579 posts and total profane words used are 3,830. This yields a percentage of word distribution of about 29.56%. This shows that profane words used by harassers are more than one quarter of the overall tokens. For the victim role, 152 profane words are used over 6,261 tokens contained in 1,355 posts, which yields a percentage of word distribution of about 2.43%. On the other hand, the percentage for the bystander_defender role is about 19.88%. This is calculated from a 730 number of profane words from 3,672 numbers of tokens contained in 424 posts. For bystander_assistant, the usage of profane words is 21 from an 88 number of tokens contained in 24 posts. This makes the percentage of profane word distribution for this role about 23.86%.

Fig. 2 presents findings of the top 10 most common profane words in four different roles in cyberbullying incidents. The graph incorporates the information about total numbers of profane words used by roles (x-axis). In the y-axis is shown the top 10 profane words used by each role in cyberbullying incidents. The profane word 'fuck' is the highest word used in all four roles in cyberbullying incidents. The 'fuck' word used by harassers is about 720, 368 is used by victims, 200 by bystander_defender, and 5 by bystander_assistant, thus making the harasser ranked first compared to others.

For harasser, profane words are ranked as follows: fuck (720 profane words), bitch (274 profane words), ugly (216 profane words), hate (173 profane words), shit (131 profane words), ass (113 profane words), cunt (111 profane words), dick (111 profane words), slut (111 profane words), and faggot (104 profane words). Victim profane words in rank are fuck (368 profane words), shit (92 profane words), bitch (88 profane words), cunt (71 profane words), dick (40 profane words), stupid (34 profane words), ass (31 profane words), pussy (28 profane words), fat (27 profane words), and dumb (22 profane words). For bystander_defender, profane words used in rank are fuck (200 profane words), shit (59 profane words), bitch (44 profane words), cunt (43 profane words), fat (32 profane words), ass (22 profane words), slut (21 profane words), stupid (18 profane words), stfu (18 profane words), and twat (16 profane words). Hence, bystander_assistants also used profane words in their cyberbullying incidents even though the total numbers for each word are less compared to other roles. The rank for bystander_assistant profane words are fuck (5 profane words), cunt (3 profane words), slut (2 profane words), lmfao (2 profane words), piss (1 profane word), bitch (1 profane word), fag (1 profane word), boob (1 profane word), damn (1 profane word), and hole (1 profane word).

We present Fig. 3 by summarizing Fig 2, which shows five highly used profane words in four roles in cyberbullying incidents. The x-axis contains information of top five profane words that are commonly used in four roles, while the y-axis displays the number of each profane words used by four roles. The total number of the 'fuck' word used by four roles is 1,293, which is the highest compared to other profane words. Secondly, the 'bitch' word used in cyberbullying incidents is 407, followed by cunt (228), slut (156), and twat (82).

5. DISCUSSION

Our focus in this work is to identify profane words used by the aforementioned roles in cyberbullying incidents using a lexicon profane word dictionary in order to analyze word distribution. This is to allow extraction of the profanity features to be implemented into the classification process of a cyberbullying detection model, and at the same time to answer the research questions. Findings revealed that each role used profane words in their con-

Table 4. Percentage of profane word distribution across four roles in cyberbullying incidents

Roles	Number of profane words	Total number of tokens	Percentage profane words distribution (%)
Harasser	3,830	12,958	29.56
Victim	152	6,261	2.43
Bystander_Defender	730	3,672	19.88
Bystander_Assistant	21	88	23.86

 https://doi.org/10.1633/JISTaP.2021.9.1.2

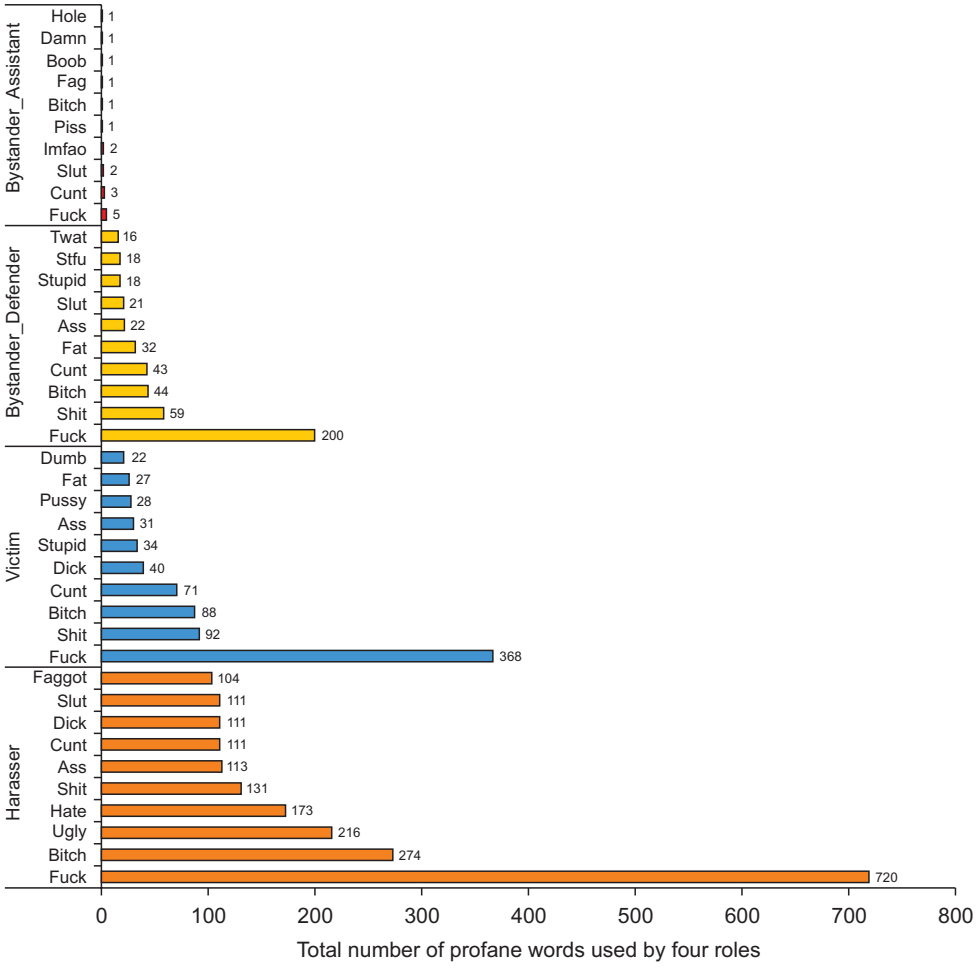
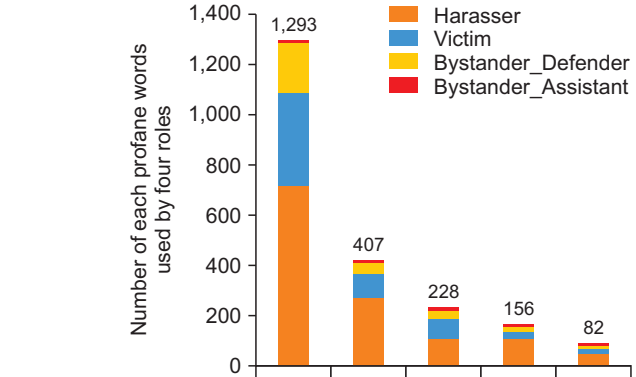


Fig. 2. Top 10 profane words distribution.



	Fuck	Bitch	Cunt	Slut	Twat
Bystander_Assistant	5	1	3	2	1
Bystander_Defender	200	44	43	21	16
Victim	368	88	71	22	17
Harasser	720	274	111	111	48

Fig. 3. Top 5 most common profane words.

version which were declared as containing a cyberbullying indication. This shows profane words can be the profanity feature in cyberbullying classification.

As mentioned, a harasser is the person who does the bullying towards victims, and profane words are always used by them to hurt others' feelings. This is to justify why the harasser is the identified role which used profane words higher than any other roles did. Meanwhile, victims may have two different reactive behaviors; either they become aggressive or become passive when being confronted by a harasser in a cyberbullying incident. Aggressive behavior appears when they try to fight back against the harasser by using the same kind of language as in the profane words, for example, "Shut the fuck up you smelly." On the other hand, a victim can react passively where he/she accepts the insults from the harasser or just replies without using any profane words, e.g., "Stop, just leave me alone." The usage of profane words by the victim could be attributed to the victim's response where the victim may

behave aggressively when confronted by the harasser; for example, "Leave Amber alone, don't be jealous and accept the fact that Amber has other friends, even if they are hipster;) so shut the fuck up u lot will never be the new Megan sweeney so positive vibes motherfuckers." Because of that, the victim is a role that also used profane words in conversations even though the percentage yielded is very low. We can see that bystander_defenders used the terms 'fuck' and 'motherfucker,' but in a positive way where they try to stop the harasser from bullying the victim, and this conforms to its definition whereby a bystander_assistant supports the harasser, likely with similar aggressive behavior to the harasser (Van Hee et al., 2018). Furthermore, it is observed that the posts by the bystander_assistant are similar to that of the harasser. This issue makes cyberbullying detection in online social networks a difficult task, especially in the role determination stage. As a whole, we can declare that the first research question is answered.

As discussed in Teh et al. (2018), the term 'fuck' has the highest score compared to other terms in year 2017. Thus, we can conclude that this kind of profane word has long been used before 2017, as this corpus was collected in 2013. Besides this, the terms 'bitch' and 'cunt' are also in the top 10 most common words used in every role in cyberbullying incidents. However, the total numbers of each profane word for each role are much different. This is caused by an imbalance of data under cyberbullying posts as mentioned in the sub-section "Data Preparation." Cyberbullying posts annotated as harrasser, victim, bystander_assistant, and bystander_defender roles have different number gaps as described in Table 2. Besides this, all of these terms are categorized as profane words and could be one of the features in a cyberbullying detection model, yet their use does not necessarily indicate the occurrence of cyberbullying; as an example, "I'm tired as fuck and have to work tomorrow." This is concluded to mean that the second research question is answered. Then, there are words used by users by substituting the original letter with another letter. For example, the letter 'u' is substituted by letter 'v' in order to convert the word 'fuck' to 'fvck,' and letter 'b' is substituted by 'v' to convert a word from 'bitch' to 'vitch.' Other than that, there are also misspelled words such as 'fck' where the original word is 'fuck.' When using a profanity dictionary to identify profane words, misspelled words cannot be identified and are considered as non-profane words. This indicates that there is a probability for a classifier to mistakenly detect cyberbullying if it only depends on profanity as a feature. However, with the information provided by Fig. 3, we can consider word embedding to improve the efficiency of the classifier for cyberbullying detection.

6. CONCLUSION AND FUTURE WORK

Recently, cyberbullying has been acknowledged as a serious national health issue in the midst of online social network users. One of the ways to solve this problem is by developing a model to detect cyberbullying in social media. Thus, this paper focuses on extracting profane words from a state-of-the-art corpus with four different roles in order to analyze profane word distribution. From the results, all of the four roles used profane words in their conversation and we concluded that users tend to use them when they are involved in cyberbullying. However, the use of profane words does not necessarily indicate cyberbullying incidents. This issue could be overcome by looking at the underlying weightage and distribution; thus, word embedding for text representation can be considered. Also, we can conclude that profane words can be one of the features used in order to develop a cyberbullying model but with the presence of other features to further improve the accuracy of the model.

In our future work, we will extract and incorporate other features to be integrated with the profanity feature in the classification process. Other than that, oversampling technique can be implemented in this corpus for a balance dataset before undergoing the classification process. Imbalance data may affect the classifier performance in training data to detect cyberbullying.

Lexical variation will also be explored to increase efficiency in the classification process of cyberbullying detection. The distribution of profane words in this result can be used in the future for an early cyberbullying detection model.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63, 433-443. https://doi.org/10.1016/j.chb.2016.05.051

Ashktorab, Z., Golbeck, J., Haber, E., & Vitak, J. (2017, June

25-28). Beyond cyberbullying: Self-disclosure, harm and social support on ASKfm. In P. Fox, D. McGuinness, & L. Poirer (Eds.), *WebSci '17: Proceedings of the 2017 ACM Web Science Conference* (pp. 3-12). ACM. https://doi.org/10.1145/3091478.3091499

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior,* 31, 259-271. https://doi.org/10.1016/j.chb.2013.10.036

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210-230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017, June 25-28). Mean birds: Detecting aggression and bullying on Twitter. In P. Fox, D. McGuinness, & L. Poirer (Eds.), *WebSci '17: Proceedings of the 2017 ACM Web Science Conference* (pp. 13-22). ACM. https://doi.org/10.1145/3091478.3091487

Computational Linguistics Research Group. (2019). *Bratreader*. https://github.com/clips/bratreader

Dinakar, K., Reichart, R., & Lieberman, H. (2011, July 17-21). Modeling the detection of textual cyberbullying. In N. Nicolov & J. G. Shanahan (Eds.), *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 11-17). AAAI Press.

Giumetti, G. W., McKibben, E. S., Hatfield, A. L., Schroeder, A. N., & Kowalski, R. M. (2012). Cyber incivility @ work: The new age of interpersonal deviance. *Cyberpsychology, Behavior and Social Networking*, 15(3), 148-154. https://doi.org/10.1089/cyber.2011.0336

Hinduja, S., & Patchin, J. W. (2018). *Cyberbullying: Identification, prevention, and response*. https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2018.pdf.

Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013, May 2-4). Detecting cyberbullying: Query terms and techniques. In H. Davis, H. Halpin, & A. Pentland (Eds.), *WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference* (pp. 195-204). ACM. https://doi.org/10.1145/2464464.2464499

Laboreiro, G., & Oliveira, E. (2014, October 6-8). What we can learn from looking at profanity. In A. Baptista, N. Mamede, S. Candeias, I. Paraboni, T. A. S. Pardo, & M. G. V. Nunes (Eds.), *Computational Processing of the Portuguese Language: 11th International Conference, PROPOR 2014. Proceedings* (pp. 108-113). Springer. https://doi.org/10.1007/978-3-319-09761-9

Moxey, N., & Bussey, K. (2020). Styles of bystander intervention in cyberbullying incidents. *International Journal of Bullying Prevention*, 2(1), 6-15. https://doi.org/10.1007/s42380-019-00039-1

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on Twitter. *Future Internet*, 12(11), 187. https://doi.org/10.3390/fi12110187

Nandhini, B. S., & Sheeba, J. I. (2015, March 6-7). Cyberbullying detection and classification using information retrieval algorithm. In S. A. Khadar (Ed.), *ICARCSET '15: Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)* (pp. 1-5). ACM. https://doi.org/10.1145/2743065.2743085

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. The *Journal of Machine Learning Research*, 12, 2825-2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December 18-21). Using machine learning to detect cyberbullying. In X. Chen, T. Dillon, H. Ishbuchi, J. Pei, H. Wang, & M. A. Wani (Eds.), *Proceedings of the 10th International Conference on Machine Learning and Applications (ICMLA 2011)* (pp. 241-244). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICMLA.2011.152

Salawu, S., He, Y., & Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1), 3-24. https://doi.org/10.1109/TAFFC.2017.2761757

Salmivalli, C. (1999). Participant role approach to school bullying: Implications for interventions. *Journal of Adolescence*, 22(4), 453-459. https://doi.org/10.1006/jado.1999.0239

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior,* 15(2), 112-120. https://doi.org/10.1016/j.avb.2009.08.007

Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior*, 22(1), 1-15. https://doi.org/10.1002/(SICI)1098-2337(1996)22:1%3C1::AID-AB1%3E3.0.CO;2-T

Sekharan, S. C., Vadivu, G., & Rao, M. V. (2018). A comprehensive study on sarcasm detection techniques in sentiment analysis. *International Journal of Pure and Applied Math-*

ematics, 118(22), 433-442. https://acadpubl.eu/hub/2018-118-22/articles/22a/63.pdf

Sheeba, J. I., Devaneyan, S. P., & Cadiravane, R. (2019). Identification and classification of cyberbully incidents using bystander intervention model. *International Journal of Recent Technology and Engineering (IJRTE),* 8(2S4), 1-6. https://doi.org/10.35940/ijrte.B1001.0782S419

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012, April 23-27). BRAT: A web-based tool for NLP-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102-107). ACL. https://dl.acm.org/doi/10.5555/2380921.2380942

Sutton, B. B. (2011). Cyberbullying: *An interview with Parry Aftab*. https://etcjournal.com/2011/02/17/7299

Talpur, B. A., & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PLoS One*, 15(10), e0240924. https://doi.org/10.1371/journal.pone.0240924

Tarmizi, N., Saee, S., & Ibrahim, D. H. A. (2020). Detecting the usage of vulgar words in cyberbully activities from Twitter. *International Journal on Advanced Science, Engineering and Information Technology*, 10(3), 1117-1122. http://doi.org/10.18517/ijaseit.10.3.10645

Teh, P. L., Cheng, C., & Chee, W. (2018, March 23-25). Identifying and categorising profane words in hate speech. In A. Gokhale & S. Zhang (Eds.), *Proceedings of 2018 the 2nd International Conference on Compute and Data Analysis* (pp. 65-69). ACM. https://doi.org/10.1145/3193077.3193078

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS One*, 13(10), e0203794. https://doi.org/10.1371/journal.pone.0203794

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014, February 15-19). Cursing in English on twitter. In S. Fussell & W.e Lutters (Eds.), *CSCW'14: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 415-425). ACM. https://doi.org/10.1145/2531602.2531734

Xu, J.-M., Jun, K.-S., Zhu, X., & Bellmore, A. (2012, June 3-8). Learning from bullying traces in social media. In J. Chu-Carroll (Ed.), *NAACL HLT '12: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 656-666). ACL. https://dl.acm.org/doi/10.5555/2382029.2382139

Zainudin, N. M., Zainal, K. H., Hasbullah, N. A., Wahab, N. A., & Ramli, S. (2016, May 16-17). A review on cyberbullying in Malaysia from digital forensic perspective. *ICICTM '16: Proceedings of 1st International Conference on Information and Communication Technology* (pp. 246-250). IEEE. https://doi.org/10.1109/ICICTM.2016.7890808

Zois, D., Kapodistria, A., Yao, M., & Chelmis, C. (2018, April 15-20). Optimal online cyberbullying detection. In M. Hayes & H. Ko (Eds.), *Proceedings of 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 2017-2021). IEEE. https://doi.org/10.1109/ICASSP.2018.8462092