Hindawi
Security and Communication Networks
Volume 2021, Article ID 5594175, 12 pages
https://doi.org/10.1155/2021/5594175

Research Article

Detection of Harassment Type of Cyberbullying: A Dictionary of Approach Words and Its Impact

Syed Mahbub, Eric Pardede, and A. S. M. Kayes

Department of Computer Science and Information Technology, La Trobe University, Melbourne VIC 3083, Australia

Correspondence should be addressed to Syed Mahbub; s.mahbub@latrobe.edu.au

Received 16 February 2021; Revised 4 May 2021; Accepted 27 May 2021; Published 4 June 2021

Academic Editor: Shehzad Ashraf Chaudhry

The purpose of this paper is to analyse the effects of predatory approach words in the detection of cyberbullying and to propose a mechanism of generating a dictionary of such approach words. The research incorporates analysis of chat logs from convicted felons, to generate a dictionary of sexual approach words. By analysing data across multiple social networks, the study demonstrates the usefulness of such a dictionary of approach words in detection of online predatory behaviour through machine learning algorithms. It also shows the difference between the nature of contents across specific social network platforms. The proposed solution to detect cyberbullying and the domain of approach words are scalable to fit real-life social media, which can have a positive impact on the overall health of online social networks. Different types of cyberbullying have different characteristics. However, existing cyberbullying detection works are not targeted towards any of these specific types. This research is tailored to focus on sexual harassment type of cyberbullying and proposes a novel dictionary of approach words. Since cyberbullying is a growing threat to the mental health and intellectual development of adolescents in the society, models targeted towards the detection of specific type of online bullying or predation should be encouraged among social network researchers.

1. Introduction

The recent widespread nature of cyberbullying has increased the importance of its detection. According to a study [1], almost 43% of the teens in the United States alone were reported to be the victims of cyberbullying at some point in time. A more recent study [2] shows the percentage has since increased to be around 59% in the United States. Cyberbullying has the same, if not a greater, negative impact on the victims in comparison to traditional bullying, as the predators usually attack a victim in relation to aspects that a person cannot change (i.e., skin colour, physical appearance, religion, and ethnicity), leaving a deeper and longer lasting impact on the victim. Sometimes the associated humiliation is enough to push the victims towards self-infliction of harm or suicide. Research [3] shows that suicidal ideation tends to increase among adolescents due to exposure to different forms of cyberbullying. Even when preventive measures are taken, the rehabilitation of victims of cyberbullying cases is a challenge for the family and society. Isolation, hypersensitivity, and self-hate dominate over the socialization process which leads to unhappy and troubled adults. Moreover, this psychological imbalance can itself create future bullies [4].

Among several challenges that make the detection of cyberbullying in OSN more difficult, present state-of-the-art solutions to cyberbullying detection do not specify the scope of bullying type in their detection model. Given the diverse types of cyberbullying that can occur on the web, it is not feasible to assume that the same detection model will be efficient in detecting every type of bullying.

In order to address this limitation of the current detection techniques, we propose a model that not only performs textual analysis as a base for training a learning model but also generates a dictionary of approach words to identify sexual harassment types of cyberbullying more accurately. The positive implications of such a detection

methodology on the overall health of online environments are prominent.

1.1. Motivation and Research Scope. The domains of social sciences and psychology have been investigating the problem of traditional bullying and cyberbullying for a long time now. A number of studies were devoted towards understanding the problem more thoroughly and some of them categorized the general term "cyberbullying" into specific types [5–8]. According to these literatures, cyberbullying can be categorized into several types, including flaming, harassment, flooding, masquerade, impersonation, cyberstalking, denigration, outing, and stalking.

Each of these types of cyberbullying has some unique traits. Our research is partly motivated from this classification and we are particularly interested in harassment type of cyberbullying, where the bully sends offensive messages to the victim. We want to narrow down our detection scope to this particular type as a general detection approach might not be appropriate for all types of bullying. We further focus our attention to one type of harassment, formally known as sexual harassment for the threat it poses and the prevalence it has among the general population. Sexual harassment is defined by Australian Human Rights Commission [9] as "any unwanted or unwelcome sexual behaviour, which makes a person feel offended, humiliated, or intimidated." When this phenomenon happens online, it takes the form of online harassment and the phenomenon is on the rise in recent years. According to research [10], 25% of American women between the ages of 18 and 24 have been the target of online sexual harassment at some point in time. The study further demonstrates that around 28% of the time, online harassment victims have found the experience extremely upsetting. These sorts of feelings, if not identified and treated straight away, can lead to severe depression and promote suicidal thoughts [3, 11].

Among the research works which have focused on the detection of online sexual predation and harassment, McGhee et al. [12] discussed the concept of different stages of predation. One of the most critical and earlier stages of such harassment is named as approach stage by the authors. This is the stage where the predator approaches the victim with some sort of sexual indication. The authors also introduced the concept of approach verbs or approach words, which are indicative words that an approach is being initiated by the predator. However, how to generate a dictionary of such approach words and how effective a generated dictionary can be in detecting cyberbullying across different types of communication channel were not addressed by previous research.

Based on the above analogy, we answer the following research questions in this paper:

> RQ1. How effective approach words can be in detecting sexual harassment type of cyberbullying across multiple social media platforms?
>
> RQ2. How can we generate a dictionary of approach words?

Our research investigates the contents of online social networks (OSNs) to answer the above research questions. The investigation leads to a number of novel research contributions towards effective detection of sexual harassment type of cyberbullying. Firstly, we propose an algorithm to generate a dictionary of approach words. Although the concept of approach stage of predation has been established in previous research works, to the best of our knowledge, none of the previous research work proposed a mechanism to generate the collection, i.e., dictionary of such approach words. Our proposed algorithm makes use of publicly available information base and generates a dictionary of approach words, which can be used as a reference for sexual harassment detection. Furthermore, we demonstrate the effectiveness of the generated dictionary by our algorithm through experimentations on datasets based on multiple OSNs. Our selection of multiple OSNs further contributes to the discussion on the difference in communication styles in such OSN platforms. Our analysis reveals how the communication style of an OSN platform can make it more or less susceptible to sexual harassment type of cybercrime.

It is imperative for our readers to understand the scope of our research in the context of text information systems. The analysis of social network data in textual format and retrieving knowledge from the data must adhere to the conceptual framework of text information systems. According to Zhai and Massung [13], the framework of a textual information system should have three areas of concerns, i.e., information access, information organization, and knowledge acquisition. Our research contributes to the area of knowledge acquisition by generating a dictionary of approach words, which can directly be used as the reference when transforming textual data into binary feature space. Figure 1 clarifies the scope of our research in the context of a text information system.

2. Materials and Methods

2.1. Dictionary of Approach Words. In order to answer our first research question (RQ1), we need to define approach words first. According to the Oxford dictionary, the word "approach" is alternatively defined as "behaviour intended to propose personal or sexual relations with someone." Based on this definition we define approach words, within the scope of cyberbullying and harassment as follows:

"Approach words are nouns, noun-phrases, verbs or verb-phrases in an online text that may reflect a subtle intension of the author towards a sexual relationship with the receiver and may eventually lead to online sexual harassment."

Following this definition, we address our research question of generating a dictionary of such words (RQ2). The question can be answered by identifying a prospective source and applying our algorithm to extract such words from that source. A competent dictionary generation requires enough examples of instances of actual sexual harassment. As a first step, we investigated several data corpus to identify a candidate that can be used as a basis for sexual harassment incidents. We identified the Perverted-Justice website (http://www.perverted-justice.com), which hosts numerous transcripts of chat-based

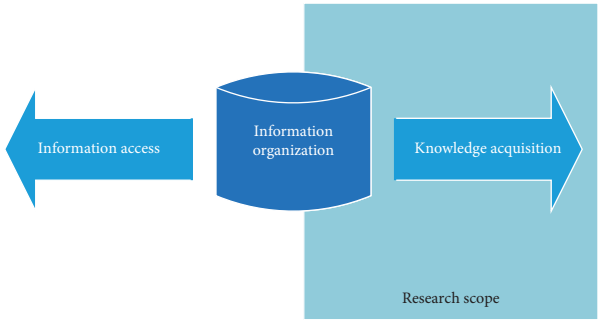

Figure 1: Scope of our research within the text information system architecture.

conversations between decoys posed as teens and convicted sexual predators in the United States. The website was first used by the research of McGhee et al. [12] and later, the large corpus of convicted transcripts was made public by the research group on ChatCoder website (http://chatcoder.com/). This website hosts a number of datasets from a larger research group and refers to some of the additional notable works [14–19] related to cyberbullying and predation detection.

The reasons we selected Perverted-Justice dataset to be the source bag-of-words for our dictionary generation algorithm are many folds. Firstly, this is the only dataset that maintains a collection of conversations that are real-life sexual harassment or predation cases. Secondly, based on these conversations, real-life predators have been lawfully convicted in the United States, which indicates the words in these conversations will include words of sexual approach. Furthermore, researchers in the domain of social and behavioural sciences relate certain human characteristics (i.e., the dark triads; Narcissism, Machiavellianism, and Psychopathy) with cyberbullying [20–23]. The same characteristics have also been identified to be related to forceful sexual predatory behaviours [24, 25]. Consequently, individuals who are convicted for online sexual harassments are more likely to possess these human characteristics, which makes them interesting subjects for studies that intersect both cyberbullying and sexual harassment. As our research falls right in that intersection, online contents generated by these individuals are of utmost relevance to us.

The corpus of Perverted-Justice chat logs hosted in ChatCoder contains 56 transcripts of various sizes. Each XML format transcript lists conversation between a unique predator and a decoy victim. We made use of Java XML SAX (Simple API for XML) parser to extract the textual contents of each of the conversations. The next step involved analysing the text and extraction of nouns, noun-phrases, verbs, and verb-phrases from the text using the Stanford CoreNLP library for Java. Although the approach words often tend to be among the parts-of-speech (POS) groups of nouns, noun-phrases, verbs, and verb-phrases, not all words that falls under this category of POS are necessarily words which might indicate sexual approach. In order to address this phenomenon, we implemented a term frequency-inverse document frequency (TF-IDF) filter to identify the rarity and normality of each word in terms of the corpus and individual chat logs.

Equations (1)–(3) below show the calculation granularity for identifying TF-IDF value for each word in the collection of words.

$$\text{TF} - \text{IDF}(t, d, D) = \text{tf}(t, d)) * \text{idf}(t, D), \quad (1)$$

where tf (t, d) is the term frequency, and idf (t, D) is the inverse document frequency.

$$\text{tf}(t, d) = \frac{n}{N}, \quad (2)$$

where N is the total number of words in a message, and n is the number of times a specific word appears in the message.

$$\text{idf}(t, D) = \log\left(\frac{N}{n}\right), \quad (3)$$

where N is the number of messages in the entire corpus, and n is the number of messages that contain the term t.

The higher the TF-IDF value, the rarer the term in the corpus. After careful observation of a range of TF-IDF values for terms that are highly profane, a threshold was selected. Any word with a TF-IDF value above the threshold was considered rare and was removed from the collection.

The dictionary was further refined by removing non-profane words and named entities such as a person's name (e.g., Alice) or a city's name (e.g., Melbourne), using Named Entity Recognition (NER) libraries from Stanford CoreNLP. Finally, the algorithm outputs an approach words dictionary of 643 words. Figure 2 illustrates a flowchart representation of the dictionary generation algorithm.

2.2. Research Workflow. Our research uses OSN posts to perform textual analysis and machine learning to identify cyberbullying posts. The steps involved in the research process are data extraction, preprocessing, experimentation, and evaluation of the learned model. As far as the OSN data is concerned, our research uses data from two different OSNs. Our primary source of data is a corpus from FormSpring, and secondary source of data is a comment corpus of YouTube.

Before we describe these datasets in more detail in the subsequent sections, we would like to take this opportunity to clarify our choice of datasets. Our goal was to select two datasets from two OSNs that differ in terms of prevalence of cyberbullying/harassment incidents and communication style. Being an open-ended communication forum, FormSpring was a strong candidate, as an OSN that is prone to cyberbullying and harassment, and at the same time, it has a specific question-answer-based communication style. Moreover, YouTube, being a topic specific video sharing forum, was a strong candidate for a secondary source of data, as an OSN that is less prone to cyberbullying/harassment, and at the same time, it has a comment-based communication style.

2.3. Description of FormSpring Dataset. The FormSpring labelled dataset was adopted from the ChatCoder website, which was used by Renolds et al. [14]. The dataset

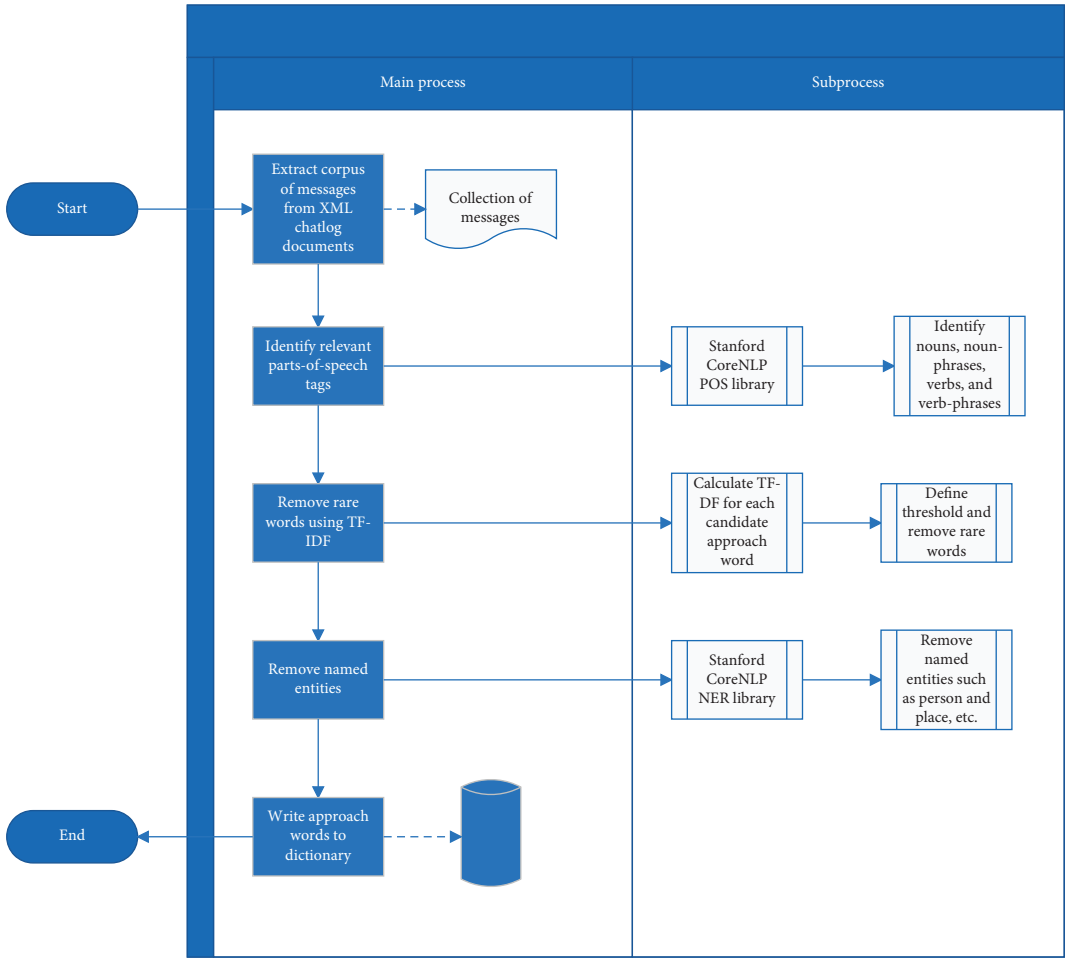

Figure 2: Algorithm for approach words dictionary generation. The figure uses a flowchart to illustrate the overall working flow of the algorithm that generated the dictionary of approach words.

represented 50 FormSpring ids along with profile information and posts associated with those ids. FormSpring is a question-and-answer-based social network. Hence, the posts contained texts in the form of questions and answers. Each post in this dataset is labelled. The labelling process is described in previous research work [14]. A "Yes" label indicates if the post contains any cyberbullying content; "No" label indicates otherwise.

2.4. Description of YouTube Dataset. The YouTube dataset consists of a total of 3,146 comments made on YouTube videos, which were extracted using YouTube Data API, a set of APIs provided by Google.

The human labelling of the corpus was done by a group of three human annotators working as freelancers and conflicting instances were assigned labels through the majority voting technique. The human annotators were instructed to label each comment as "Yes" (if the comment is an instance of cyberbullying) or "No" (if the comment is innocent), based on the following rule-of-thumb: a comment gets labelled "Yes" if it was made with an intention of:

(i) Bullying someone

(ii) Humiliating someone

(iii) Approaching someone with sexual motive

Joint probability of agreement was selected as a measure of interrater reliability among the three annotators and 88.05% of the time they agreed on the assigned class label.

2.5. Feature Space Design. Based on our study of the previous literature and focused intent, our proposed methodology includes the analysis of the following textual attributes to be considered for training a learner model:

(1) Binary representation of whether a word is present in the text within a list of swear words (e.g., arsehole and faggot)

(2) Binary representation of whether a word is present in the text within a list of malevolent words (e.g., adverse and banal)

(3) Presence of negative words in front of swear or malevolent words to neutralize the negative meaning (e.g., "not an" in front of arsehole)

(4) Presence of positive and negative emoticons or smileys

(5) Binary representation of whether a word is present in the text within a list of approach words (e.g., words indicating sexual approach)

The approach word list was generated by our proposed algorithm. The lists of swear words and malevolent words were downloaded from Noswearing website. A few examples of approach words and other binary features along with the labels assigned to each of the comments are provided in Table 1 for a clear understanding of how the raw comments were treated and labelled. The examples are extracted from the YouTube corpus.

Figure 3 below shows a graphical representation of the proposed workflow, including the context of approach word dictionary usage and binary feature vector generation. The collection of feature vectors for the corpus is then used as the training data in the machine learning phase of the experiment.

The machine learning tool which is used in this work is Waikato Environment for Knowledge Analysis (WEKA) and the evaluation is done based on several mining metrics, i.e., accuracy, precision, and recall. Since identifying bullying instances is a classification problem, for both datasets, we performed multiple experiments using J48 decision tree classifier, JRip rule-based classifier, and Naïve Bayes classifier, available in the WEKA suite. Our methodology also included oversampling of the positive instances of cyberbullying during the data preprocessing step in order to correct the bias towards innocent comments. A 10-fold cross validation was performed to reduce the chances of overfitting.

3. Results and Discussion

3.1. Experimental Setup for FormSpring and YouTube Datasets. For each of the classification algorithms, we conducted two experiments (let's label them *A* & *B*). For experiment *A*, the presence of approach words was not included in the feature space, whereas for experiment *B*, the binary feature vector was generated using a feature space that contained the presence of approach words. These experiments were designed to identify the effect of the approach words in the feature space on the overall model performance. The FormSpring dataset initially contained 6.88% of positive cyberbullying instances and an oversampling of factor 5 increased the percentage of positive cyberbullying instances to 34.4%.

Similar experiments were conducted on the YouTube dataset. Experiment *C* did not include the approach words in the feature space, whereas experiment *D* did. The YouTube dataset contained 4.32% of positive cyberbullying instances and an oversampling of factor 5 increased the percentage to 21.44%.

3.2. Summary of Results. Table 2 lists the accuracy (a), precision (p), and recall (r) for class label "Yes", for all four experiments (*A*, *B*, *C*, & *D*) for the three classification algorithms used.

As stated earlier, the performance measures across three classifiers are consistent for both datasets. However, inclusion of approach words in the feature space did not improve the performance measures for YouTube but increased the performance measures for FormSpring. We analyse the underlying reasons for that in the next section, along with certain differences in the nature of data for these two platforms.

4. Discussion

Before we begin the discussion on analysis, we want to emphasize a couple of points. Firstly, the rationale behind selecting multiple machine learning algorithms was not to identify which one performs better for a certain type of dataset, rather the rationale was to ensure the consistency of results across different algorithms. Moreover, several recent research works in the field of cyberbullying detection in general made use of deep learning algorithms to train detection models. However, the focus of our research was not to propose a novel machine learning algorithm, rather the focus was to propose an algorithm that can generate a dictionary of approach words and demonstrate the effectiveness of the generated dictionary. We were able to demonstrate the effectiveness of the dictionary by setting up multiple experimental feature spaces with and without certain features. With our approach of feature space design, traditional machine learning algorithms were deemed sufficient. Secondly, due to the difference in dataset and feature space, we did not deem the comparison of our performance measures with previously published work of similar nature, a logical one. Having said that, the emphasis of our experimental setup was to investigate whether the proposed dictionary of approach words makes a difference in detecting sexual harassment type of cyberbullying.

As we can see in Table 2, experiments *A* & *B* were conducted using the FormSpring dataset with and without the approach words, respectively. Experiment *B*, which included the approach words in the feature space, showed improvement in terms of performance measures over experiment *A*, which did not include the approach words in the feature space. The improvements were consistent across all three classifiers. For J48 decision tree, the accuracy of the model increased to 81.60% in experiment *B* from 80.07% in experiment *A*. The recall value for class label "Yes" remained at 0.54 across the two experiments. However, the precision value for class label "Yes" increased from 0.80 in experiment *A* to 0.86 in experiment *B*. The respective measures for JRip and Naïve Bayes algorithms were similar to J48. Unlike experiments *A* & *B*, where inclusion of approach words in the feature space increased the performance measures, experiments *C* & *D* showed no change in accuracy, precision, and recall for J48 decision tree classifier. For J48 decision tree classifier, the accuracy remained unchanged at 89.56% from experiment *C* to experiment *D*. The values of precision and recall for class label "Yes" remained unchanged across experiments *C* and *D* at 0.74 and 0.81, respectively. The measures for experiment *C* & *D* for JRip and Naïve Bayes

Table 1: Examples of sample comments with features and labels.

Comment	Feature present	Labels assigned by human annotators			Final label
		Annotator 1	Annotator 2	Annotator 3	
Denilson Igwe is very stupid.	Malevolent word (stupid)	Yes	Yes	Yes	Yes
He has the gay sass to him.	Swear word (gay)	Yes	Yes	No	Yes
She is a lot of things but not a bitch.	Negative word before swear word	No	No	No	No
Do you want to meet in a hotel room?	Approach word "hotel" (in a specific context)	Yes	No	Yes	Yes
Are you alone in the house?	Approach word "alone" (in a specific context)	Yes	Yes	Yes	Yes

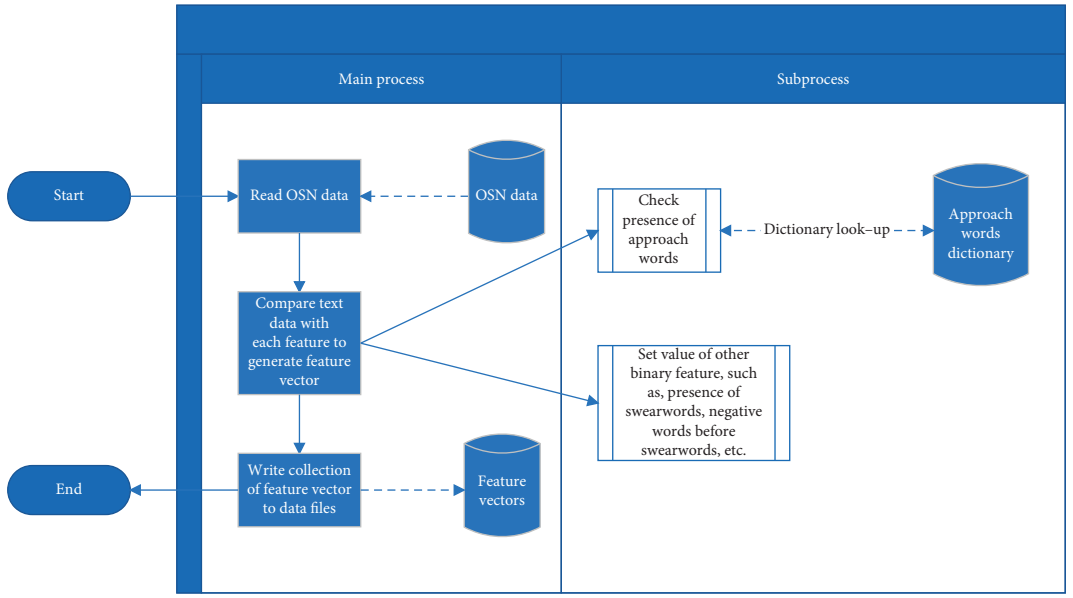

Figure 3: The use of approach words dictionary within the context of feature vector generation process.

Table 2: Summary of performance measures for different experiments.

Training data	FormSpring dataset						YouTube dataset					
Feature space	Without approach words (A)			With approach words (B)			Without approach words (C)			With approach words (D)		
Algorithm	a (%)	p	r	a (%)	p	r	a (%)	p	r	a (%)	p	r
J48	80.07	0.80	0.54	81.60	0.86	0.54	89.56	0.74	0.81	89.56	0.74	0.81
JRip	80.07	0.80	0.54	81.60	0.86	0.54	89.86	0.75	0.81	89.79	0.74	0.81
Naïve Bayes	80.05	0.80	0.55	81.60	0.86	0.55	89.56	0.74	0.81	89.56	0.74	0.81

classifiers were also very similar to the measures reported for J48 decision tree.

The performance measures in Table 2 further reveal that the accuracy, precision, and recall of the YouTube dataset were significantly higher than similar experiments conducted with FormSpring dataset. The reason of such differences lies within the nature of the data for these platforms. The comments on YouTube are longer and seldom part of a conversation between commenters whereas the FormSpring text in the form of question-answer is mostly short-lived and can be perceived as casual conversations between the platform users. Often, such short-lived, casual conversations need to be analysed as a whole, rather than individual instances of texts. For these types of platforms, window of communication can reveal further information. Then again, our analysis revealed that approach words do not necessarily improve cyberbullying detection on platforms like YouTube, where the comments are subject specific, and the scope is narrow. The categories of videos extracted from YouTube, such as Education, Entertainment, People & Blogs, etc., also suggest why the comments are less prone to sexual harassment type of bullying and more specific to opinion expressions on video content. However, inclusion of approach words in the feature space improves the performance measures in detection of cyberbullying on FormSpring data, which suggest that the platform is more prone to such type of harassments.

Table 3: Instance-based analysis snapshot.

Part of conversation in FormSpring	Labels		
	Actual label	Model prediction (feature space with no approach words)	Model prediction (feature space with approach words)
Are you home alone?	Yes	No	Yes
Can we meet in a hotel room alone?	Yes	No	Yes
Can you send me your pics?	Yes	No	Yes
Do you have a bf? Can you consider me instead?	Yes	No	Yes

Table 4: Summary of related works.

Research work	Publication year	Source of data	Content analysis	Contextual analysis	Learning algorithm class	Reference word list sources	Dictionary/ word list generation proposal
Agrewal and Awekar [26]	2018	FormSpring, Twitter, Wikipedia	Yes	No	Deep neural network, support vector machine, logistic regression, random forest, Naïve Bayes	None	No
Aind et al. [27]	2020	Multiple publicly available datasets	Yes	No	Novel algorithm based on reinforcement learning	GitHub reference wordlists for profanity and sentiment dictionaries (refer to paper)	No
Balakrishnan et al. [28]	2020	Twitter	Yes	Yes	Random forest, decision tree	Unspecified	No
Banerjee et al. [29]	2019	Twitter	Yes	No	Deep neural network	None	No
Cheng et al. [30]	2019	FormSpring, Twitter	Yes	No	Random forest, Extratree, AdaBoost	Unspecified	No
Cheng et al. [31]	2019	Instagram	Yes	Yes	Novel algorithm (hierarchical attention networks for cyberbullying detection)	Unspecified	No
Dadvar et al. [32]	2013	YouTube	Yes	Yes	Support vector machine	Noswearing website	No
Dani et al. [33]	2017	Twitter, MySpace	Yes	Yes	Linear regression, sparse learning, support vector machine	Unspecified	No
Dinakar et al. [34]	2011	YouTube	Yes	No	Naïve Bayes, rule-based JRip, decision tree, support vector machine	Unspecified	No
Hosseinmardi et al. [35]	2015	Instagram	Yes	Yes	Statistical analysis	Unspecified	No
Hosseinmardi et al. [36]	2014	Instagram, Ask.fm	Yes	Yes	Statistical analysis	Unspecified	No

Table 4: Continued.

Research work	Publication year	Source of data	Content analysis	Contextual analysis	Learning algorithm class	Reference word list sources	Dictionary/ word list generation proposal
Iwendi et al. [37]	2020	Kaggle dataset from Facebook, Twitter, Instagram	Yes	No	Deep learning models	None	No
Kontostathis [17]	2009	Perverted-Justice	Yes	No	Decision tree, K-mean clustering	Predation dictionary (refer to paper)	No
Kontostathis et al. [18]	2012	Perverted-Justice	Yes	No	Decision tree, rule-based classifier	Predation dictionary (refer to paper)	No
Kontostathis et al. [19]	2013	FormSpring	Yes	Yes	Essential dimensions for LSI	Noswearing website	No
Lu et al. [38]	2020	Chinese Weibo, Twitter	Yes	No	Convolutional neural network	None	No
McGhee et al. [12]	2011	Perverted-Justice	Yes	No	Decision tree, rule-based classifier, K-nearest neighbour	Predation dictionary (refer to paper)	No
Nahar et al. [39]	2014	MySpace, Kongregate, Slashdot	Yes	Yes	Fuzzy C-mean clustering, fuzzy support vector machine	Unspecified	No
Ptaszynski [40]	2019	Multiple unofficial school websites and forums (see paper for more information)	Yes	No	Novel brute-force pattern extraction algorithm	None	No
Rafiq et al. [41]	2018	Vine	Yes	No	AdaBoost, logistic regression, incremental classifier	None	No
Raisi and Huang [42]	2018	Twitter, Ask.fm, Instagram	Yes	Yes	Novel participant vocabulary consistency	Noswearing website	No
Renolds et al. [14]	2011	FormSpring	Yes	No	Decision tree, rule-based classifier, support vector machine, K-nearest neighbour	Noswearing website	No
Tahmasbi and Rastegari [43]	2018	Twitter	Yes	Yes	Decision tree, rule-based classifier, support vector machine, logistic regression, AdaBoost, Naïve Bayes	Unspecified	No
Van Hee et al. [44]	2018	Ask.fm	Yes	No	Support vector machine	Google profanity list	No

Table 4: Continued.

Research work	Publication year	Source of data	Content analysis	Contextual analysis	Learning algorithm class	Reference word list sources	Dictionary/ word list generation proposal
Wang et al. [45]	2020	Instagram, Vine	Yes	No	Novel multimodal cyberbullying detection framework (based on neural network)	None	No
Xu et al. [46]	2012	Twitter	Yes	Yes	Logistic regression, support vector machine, Naïve Bayes, latent topic models	None	No
Yao et al. [47]	2019	Instagram	Yes	No	Novel sequential hypothesis testing model CONciSE	Noswearing website	No
Yin et al. [15]	2009	MySpace, Kongregate, Slashdot	Yes	Yes	Support vector machine	Noswearing website	No
Zhao et al. [48]	2020	Twitter	Yes	No	Support vector machine, logistic regression, random forest, and multiple deep learning models	None	No
Zhong et al. [49]	2016	Instagram	Yes	Yes	Support vector machine, convolutional neural network, deep learning models	None	No
Gencoglu [50]	2021	Jigsaw, Twitter, WikiDetox, Gab Hate Corpus	Yes	Yes	Deep neural network	None	No
Cheng et al. [51]	2020	Instagram, Vine	Yes	Yes	Unsupervised Gaussian mixture model	Unspecified	No
Kumar and Sachdeva [52]	2021	YouTube, Instagram, Twitter	Yes	No	Convolutional neural network, deep neural network	None	No
Dadvar and Eckert [53]	2020	FormSpring, Wikipedia, Twitter, YouTube,	Yes	No	Deep neural networks	None	No
Wang et al. [54]	2020	FormSpring, Twitter	Yes	No	Word2Vec, word similarity scheme	Noswearing website	No
Fang et al. [55]	2021	Twitter, Wikipedia	Yes	No	Neural network with gated recurrent unit	None	No
Rezvani et al. [56]	2020	Instagram, Twitter	Yes	Yes	Neural network	Google profanity list	No
Current work		YouTube + FormSpring	Yes	No	Decision tree, Naïve Bayes, rule-based classifiers	Noswearing website + generated dictionary	Yes

Moreover, instead of solely depending on the performance measures of the classification models, we opted for an instance-based approach for further analysis, which looks for improved detection of true positive cases of sexual harassments in terms of individual instances. Table 3 lists several instances of the FormSpring dataset below that distinguishes the outcome of the model for two different feature spaces (i.e., feature space with approach words and feature space without approach words).

The example instances in Table 3 signify the importance of approach words in detecting subtle hints of sexual approach by a predator. If these sorts of online comments can be detected automatically, and the relevant authorities can be notified, that would help significantly in reducing the number of cyberbullying and sexual harassment cases in the online environment.

Our experimental findings directly address our research questions, which were outlined in the motivation section. Firstly, our acquired results demonstrate that platforms like FormSpring, which facilitates question-answer-based and short-lived casual conversations among users, are more prone to harassment type of bullying. As a result, the consideration of approach words improves the performance measures of cyberbullying detection model, when the model is trained and evaluated with FormSpring dataset. On the contrary, platforms such as YouTube are less prone to harassment type of cyberbullying incidents, and thus inclusion of approach words in the feature space does not necessarily improve the detection accuracy of cyberbullying when the model is trained and evaluated using the YouTube dataset (RQ1). Moreover, our algorithm for the dictionary generation of approach words, which was outlined in the methodology section, provides a clear mechanism to generate a collection of approach words from publicly available information base (RQ2).

Finally, to put our work in the context of related works in the domain of cyberbullying detection, we present an overview of several aspects of notable research works, along with the proposed approach, in Table 4. For each research work, we identify their data source, whether they have performed any contextual analysis along with the content analysis, their learning approach, and reference dictionary (if used any). Contextual analysis refers to analysis of contextual information, such as user profile, demographic information, network information, or anything that is not directly part of the content being analysed. The table further highlights the fact that our research work proposes a mechanism to generate a dictionary of approach words, which has not been proposed by any relevant work in the field of cyberbullying detection.

5. Conclusions

The majority of approaches aimed at automatic detection of cyberbullying events on social media focus on content-based analysis and propose feature spaces that can train machine learning models to detect such events. In this paper, we propose a systematic approach of such feature space design that takes the generation of keyword dictionary into consideration. We focus on the dictionary generation of approach words from real-life case studies and demonstrate their effect in multiple OSNs. Through our experimental findings, we further demonstrate the effectiveness of such a methodology to detect sexual harassment type of cyberbullying.

Due to the ever-changing nature of today's OSN, the future extension of this work can take many directions. First, the number of attributes considered in the feature space can be extended to improve the model for textual analysis. In addition to the binary representation of each of the attributes, normalized values can also be used in the future to represent the severity or weight of certain features. Network features can also be analysed in the future to calculate the popularity or activeness of a user within a local network or within the OSN. Sarcasm detection, the context of posts, and the window of communication for consecutive posts can be considered for future work as well. Structured framework such as the one proposed in previous research [57] might help the approach to be generalised across multiple platforms, where the aforementioned diverse categories of features are taken into consideration. Furthermore, for future extensions, contextual user information can also be utilised similar to other domains [58], which might add significant value. Additional perspective into the nature of bullying behaviour, that takes into account the user session specific information [59], can also be considered in future works. However, the improvement of approach words dictionary generation technique will be of prime importance within the scope of our research. The nature of predation needs to be studied further for a more compact and accurate dictionary.

Data Availability

The principal datasets used in this research can be downloaded from the ChatCoder (http://www.chatcoder.com/data.html) website.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] M. Ybarra, "Trends in technology-based sexual and non-sexual aggression over time and linkage to non-technology aggression," in *National Summit on Interpersonal Violence and Abuse across the Lifespan: Forging a Shared Agenda*-Growing Up With Media, Houston, TX, USA, 2010.

[2] Pew Research Center, "Online harassment," 2018, https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/.

[3] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of Suicide Research*, vol. 14, no. 3, pp. 206–221, 2010.

[4] A. G. Garrett, *Bullying in American Schools*, McFarland & Company Inc, Jefferson, NC, USA, 2003.

[5] N. E. Willard, *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*, Research Press, Champaign, IL. USA, 2007.

[6] D. Maher, "Cyberbullying: an ethnographic case study of one Australian upper primary school class," *Youth Studies Australia*, vol. 27, no. 4, 2008.

[7] F. Mishna, M. Saini, and S. Solomon, "Ongoing and online: children and youth's perceptions of cyber bullying," *Children and Youth Services Review*, vol. 31, no. 12, 2009.

[8] A. Chakraborty, Y. Zhang, and A. Ramesh, "Understanding types of cyberbullying in an anonymous messaging application," in *Proceedings of the Companion Web Conference 2018*, pp. 1001–1005, Lyon, France, April 2018.

[9] Australian Human Rights Commission, "Sexual harassment in the workplace—the legal definition of sexual harassment," 2019, https://humanrights.gov.au/our-work/sexual-harassment-workplace-legal-definition-sexual-harassment.

[10] Pew Research Center, "Online harassment," 2014, http://www.pewinternet.org/2014/10/22/online-harassment/.

[11] S. Hinduja and J. W. Patchin, *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying*, CORWIN, Nampa, ID, USA, 2008.

[12] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, "Learning to identify internet sexual predation," *International Journal of Electronic Commerce*, vol. 15, no. 3, 2011.

[13] C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, Association for Computing Machinery and Morgan & Claypool, New York, NY, USA, 2016.

[14] K. Renolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops*, Honolulu, HI, USA, December 2011.

[15] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," *Content Analysis in the WEB*, vol. 2, 2009.

[16] A. Leatherman, *Luring Language and Virtual Victims: Coding Cyber Predators' On-Line Communicative Behavior*, Ursinus College, Collegeville, PA, USA, 2009.

[17] A. Kontostathis, "Chatcoder: toward the tracking and categorization of internet predators," in *Proceedings of the Text Mining Workshop 2009 Held in Conjunction with the Ninth Siam International Conference on Data Mining (SDM 2009)*, Sparks, NV, USA, April 2009.

[18] A. Kontostathis, A. Garron, K. Reynolds, W. West, and L. Edwards, "Identifying predators using ChatCoder 2.0," in *Proceedings of the CLEF (Online Working Notes/Labs/Workshop)*, Rome, Italy, September 2012.

[19] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in *Proceedings of the 5th Annual Acm Web Science Conference*, pp. 195–204, Paris France, May 2013.

[20] A. K. Goodboy and M. M. Martin, "The personality profile of a cyberbully: examining the dark triad," *Computers in Human Behavior*, vol. 49, pp. 1–4, 2015.

[21] M. van Geel, A. Goemans, F. Toprak, and P. Vedder, "Which personality traits are related to traditional bullying and cyberbullying? a study with the big five, dark triad and sadism," *Personality and Individual Differences*, vol. 106, pp. 231–235, 2017.

[22] S. Pabian, C. J. S. De Backer, and H. Vandebosch, "Dark triad personality traits and adolescent cyber-aggression," *Personality and Individual Differences*, vol. 75, pp. 41–46, 2015.

[23] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on twitter using big five and dark triad features," *Personality and Individual Differences*, vol. 141, pp. 252–257, 2019.

[24] V. Zeigler-Hill, A. Besser, J. Morag, and W. Keith Campbell, "The dark triad and sexual harassment proclivity," *Personality and Individual Differences*, vol. 89, pp. 47–54, 2016.

[25] P. K. Jonason, M. Girgis, and J. Milne-Home, "The exploitive mating strategy of the dark triad traits: tests of rape-enabling attitudes," *Archives of Sexual Behavior*, vol. 46, no. 3, pp. 697–706, 2017.

[26] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proceedings of the European Conference on Information Retrieval*, March 2018.

[27] A. T. Aind, A. Ramnaney, and D. Sethia, "Q-bully: a reinforcement learning based cyberbullying detection framework," in *Proceedings of the 2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, June 2020.

[28] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using twitter users' psychological features and machine learning," *Computers & Security*, vol. 90, Article ID 101710, 2020.

[29] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of cyberbullying using deep neural network," in *Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, Coimbatore, India, March 2019.

[30] L. Cheng, R. Guo, and H. Liu, "Robust cyberbullying detection with causal interpretation," in *Proceedings of the Companion Proceedings of the 2019 World Wide Web Conference*, pp. 169–175, San Francisco, CA, USA, May 2019.

[31] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *Proceedings of the Proceedings of the 2019 SIAM International Conference on Data Mining*, Calgary, Canada, February 2019.

[32] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Proceedings of the European Conference on Information Retrieval*, Berlin, Heidelberg, March 2013.

[33] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, September 2017.

[34] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," *The Social Mobile Web*, vol. 11, no. 2, pp. 11–17, 2011.

[35] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in *Proceedings of the International Conference on Social Informatics*, December 2015.

[36] H. Hosseinmardi, S. Li, Z. Yang et al., "A comparison of common users across instagram and ask. fm to better understand cyberbullying," in *Proceedings of the IEEE Fourth International Conference on Big Data and Cloud Computing*, Sydney, Australia, December 2014.

[37] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, pp. 1–14, 2020.

[38] N. Lu, G. Wu, Z. Zhang, Y. Zheng, Y. Ren, and K.-K. R. Choo, "Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts,"

Concurrency and Computation: Practice and Experience, vol. 32, no. 23, Article ID e5627, 2020.

[39] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Proceedings of the 25th Australasian Database Conference (ADC)*, July 2014.

[40] M. Ptaszynski, P. Lempa, F. Masui et al., "Brute-force sentence pattern extortion from harmful messages for cyberbullying detection," *Journal of the Association for Information Systems*, vol. 20, no. 8, p. 4, 2019.

[41] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra, "Scalable and timely detection of cyberbullying in online social networks," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pp. 1738–1747, Pau France, April 2018.

[42] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection with participant-vocabulary consistency," *Social Network Analysis and Mining*, vol. 8, no. 1, p. 38, 2018.

[43] N. Tahmasbi and E. Rastegari, "A socio-contextual approach in automated detection of cyberbullying," in *Procesedings of the Proceedings of the 51st Hawaii International Conference on System Sciences*, Waikoloa Village, Hi, USA, January 2018.

[44] C. Van Hee, G. Jacobs, C. Emmery et al., "Automatic detection of cyberbullying in social media text," *PLoS One*, vol. 13, no. 10, Article ID e0203794, 2018.

[45] K. Wang, Q. Xiong, C. Wu, M. Gao, and Y. Yu, "Multi-modal cyberbullying detection on social networks," in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, July 2020.

[46] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 656–666, Montréal, Canada, June 2012.

[47] M. Yao, C. Chelmis, and D.-S. Zois, "Cyberbullying ends here: towards robust detection of cyberbullying in social media," in *Proceedings of the The World Wide Web Conference*, San Francisco, CA, USA, May 2019.

[48] Z. Zhao, M. Gao, F. Luo, Y. Zhang, and Q. Xiong, "LSHWE: improving similarity-based word embedding with locality sensitive hashing for cyberbullying detection," in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, July 2020.

[49] H. Zhong, H. Li, A. C. Squicciarini et al., "Content-driven detection of cyberbullying on the instagram social network," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 3952–3958, New York, NY, USA, July 2016.

[50] O. Gencoglu, "Cyberbullying detection with fairness constraints," *IEEE Internet Computing*, vol. 25, no. 1, pp. 20–29, 2021.

[51] L. Cheng, K. Shu, S. Wu, Y. Silva, D. Hall, and H. Liu, "Unsupervised cyberbullying detection via time-informed deep clustering," in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, Galway, Ireland, October 2020.

[52] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multimedia Systems*, pp. 1–10, 2021.

[53] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models," in *Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery*, pp. 245–255, Springer, Bratislava, Slovakia, September 2020.

[54] K. Wang, Y. Cui, J. Hu, Y. Zhang, W. Zhao, and L. Feng, "Cyberbullying detection, based on the fasttext and word similarity schemes," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–15, 2020.

[55] Y. Fang, S. Yang, B. Zhao, and C. Huang, "Cyberbullying detection in social networks using Bi-gru with self-attention mechanism," *Information*, vol. 12, no. 4, p. 171, 2021.

[56] N. Rezvani, A. Beheshti, and A. Tabebordbar, "Linking textual and contextual features for intelligent cyberbullying detection in social media," in *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, pp. 3–10, Chiang Mai, Thailand, November 2020.

[57] S. Mahbub, E. Pardede, A. S. M. Kayes, and W. Rahayu, "Controlling astroturfing on the internet: a survey on detection techniques and research challenges," *International Journal of Web and Grid Services*, vol. 15, no. 2, pp. 139–158, 2019.

[58] S. Mahbub and E. Pardede, "Using contextual features for online recruitment fraud detection," in *Proceedings of the 27th International Conference on Information Systems Development (ISD2018)*, Lund, Sweden, August 2018.

[59] L. Cheng, Y. N. Silva, D. Hall, and H. Liu, "Session-based cyberbullying detection: problems and challenges," *IEEE Internet Computing*, vol. 25, no. 2, pp. 66–72, 2021.