Advance Access publication on 19 April 2021
https://doi.org/10.1093/comjnl/bxab019

Detecting Cyberbullying across Social Media Platforms in Saudi Arabia Using Sentiment Analysis: A Case Study

Suliman Mohamed Fati*

College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
**Corresponding author: smfati@yahoo.com*

Twitter has become an open space for the users' interactions and discussions on diverse trending topics. One of the issues raised on social media platforms is the misunderstanding of 'freedom of speech', which in turn, leads us to a new social and behavioral attack: cyberbullying. Cyberbully affects both individuals and societies. Despite tough sanctions globally and locally, cyberbullying is still a serious issue, which needs further consideration. Thus, this research aims to address this issue by proposing a framework, based on sentiment analysis, to detect cyberbullying in the tweets stream. The proposed framework in this paper extracts the tweets from Twitter. Then, the preprocessing through tweets tokenization was applied to remove noise from tweets and also symbols and phrases such as http, emoji faces, hash tag symbols, mention symbols and retweet. After data tokenization, the proposed system classifies the tweets, based on extracted keywords from both experts and potential victims, using deep-learning classification algorithm with 70% of dataset samples used for the training purpose, and 30% of the dataset samples used for the testing purpose. The experimental results show the ability of proposed methodology to detect cyberbullying effectively with accuracy 81%.

Keywords: cyberbully; machine learning; social media; detection; Saudi Arabia

Received 14 January 2020; Revised 21 January 2021; Editorial Decision 24 February 2021
Handling editor: Sabah Mohammed

1. INTRODUCTION

With tremendous advancement in the Internet 2.0 technology, social media platforms, such as Twitter and Facebook, have become enormously prevalent and play a significant role in changing human lives [1]. Particularly, social media platforms integrated the daily activities, such as education, business, entertainment, e-government, etc., into human's daily life. Thus, social media is nowadays an integral daily life element. However, without doubt, the use of technologies, including social media, by young people might expose them to threats such as cyberbullying, which is one of the most significant social attacks happening on social media platforms nowadays. Cyberbully is defined as posting offensive messages against an individual through digital means, often anonymously [3]. The word cyberbully originated from the word bully because of the similar intentions and similar effect caused. Dictionary defines bully as an annoying person, who feels himself sturdier than others, and insults or threatens others in an offensive way [3].

The increasing number of cyberbullying cases triggers the red alarm of cyberbullying danger especially to teenagers and young adults, who are mostly inconsiderate and juvenile. At this age, they take everything seriously without knowing how to deal with social issues; this leads them to express their feelings on social media in a way that may harm others [4]. Therefore, there are a number of global initiatives to prevent cyberbully and enhance the safety of Internet users, especially the children.

Saudi Arabia, which is the largest active population of social media users in the region, also suffers from this issue. As stated by Simon Kemp in Digital 2020 report [30], Saudis are the largest active population of social media in the region, who are the top users of Twitter, Instagram and Snapchat. In Saudi Arabia, 32.23 million of the total population (33.85 million) are digitalized with the Internet access, which represent 93%

of Saudis. A total of 25 million are active social media users, which represent 72% of the population, and around 18.96 million of them are frequent Twitter users, which represents 56% of the social media [2].

The negative consequences of cyberbullying cannot be ignored; it affects the mental health of victims, causing psychological conditions such as depression, low self-esteem, disability, social anxiety, suicide, fear, poor social relations with peers, and some studies also indicate the emergence of cases and complaints of pain, headache, stomach pain, difficulty in sleeping and other physical symptoms [5]. Notably, the studies have shown that bullies often suffer from the psychological conditions, therefore they bullied someone else and made them suffer too [6]. Thus, cyberbullying is like an infection, which may produce an aggressive society, particularly, in case of university and school students, who are millennial with high technology proficiency. Thus, the need for detecting and controlling cyberbullying on the social media is crucial for individuals and society.

Therefore, this research aims to build a framework to detect cyberbullying by extracting the tweets, analyze them using sentiment analysis techniques based on predefined keywords, and then classify the tweets to cyberbully or not-cyberbully. The users can filter their Twitter account news feed based on this framework. To achieve the aim of this paper, two research questions should be answered: (1) how to identify offensive language on social media? and (2) How to detect cyberbully intentions on social media? The fact-finding techniques will be used to answer Q1, while Q2 will be answered using sentiment analysis. The details will be discussed throughout the rest of this paper. The rest of this paper is organized as follows. Section 2 is dedicated to Background and Related works, which discuss the causes and impact of cyberbullying; Section 3 presents Proposed cyberbullying detection Framework; Section 4 is dedicated to experimental results and discussion. Finally, Section 5 presents the conclusion and future work.

2. BACKGROUND AND RELATED WORKS

Due to its popularity as a dominant social media platform, Twitter has become an open space for the users' interactions and discussions on diverse trending topics. As a result, the users of Twitter, and other social media, enjoy the freedom of speech, particularly when they use fake accounts, or when they hide their identities for the sake of anonymity. Therefore, majority of social media users share their personal life matters, disclose their private information, and in many cases, they express their opinions without considering their posts could harm others due to misunderstanding of freedom of speech. Furthermore, the social media users might defend their opinion in an aggressive way, which hurt other people intentionally or unintentionally during the discussion or public dialog. Such conflict and misunderstanding of the controversial topics on the social media leads us to a new social attack, called cyberbully, which is one of the most significant social attacks happening on the social media platforms nowadays.

According to [31], there are 1.5 million cyberbullying incidents globally every day. More than 87% of young Internet users perceived cyberbullying occurred. Locally, more than 3.6 million people in Saudi Arabia were subjected to cyberbullying in 2019. This indicator is sufficient to consider the cyberbully issue. Therefore, this research aims to help in minimizing the impact of cyberbully in social media, particularly Twitter, by automating the detection process of cyberbully. To understand how cyberbullying can be detected on social media, the need to investigate slightly the cyberbullying causes, types and the social media rules to prevent cyberbullying is present. Thus, the next subsection is dedicated to explore these details.

2.1. Cyberbullying: Causes and Types

Remarkably, cyberbullying can be associated with many social factors such as lack of support from friends, proactive aggression, jealousy, preconception, embarrassment, arrogance, culpability, anger and exposure to vehemence and its justification. In addition, intolerance for disability, anonymity approval, boredom feel better, instigate jealousy, no perceived consequences projection of feelings, protection reinvention of self and revenge also cause cyberbullying [7–10]. Besides, relationship issues, such as break-ups, envy and gang-up, activate cyberbullying. The authors in [8] insist on the exposure of victims to muscularly negative effects (especially on their social well-being) with the lack of appropriate or effective reactive behaviors from both peers and schools as a cyberbullying reason.

Based on the aforementioned causes, cyberbullying can be performed in different means based on the bully's intention. The different types of cyberbullying include harassment, cyberstalking, flaming, exclusion, outing, masquerading and denigration. The following lines will discuss and summarize these types. Harassment is an unpleasant behavior wherein the bully keeps sending flood of offensive and malicious mails/messages/posts repeatedly, causing the victim feel frightened or humiliated [11]. These abusing messages are sent in a quick fire, which annoys the victims [33]. On the other side, cyberstalking is another form of harassment, whereby the bully continuously sends discourteous and hostile mails/messages/-posts [12]. Cyberstalking involves sending continuous insulting messages to the victim trying to convince him/her to respond to the bully or quit the social media. Unlike harassment, one distinct feature of cyberstalking is the possibility of relation existence between the bully and victim [34]. Another form of harassment is called masquerading wherein the bully is practicing a fake identity to feel secure of legal consequences, or impersonates an identity that is known or closer to the victim [12]. According to [34], the relation nature between the bully and victim determines the crime and its impact on the victim. Another type of cyberbully, which is considered more aggressive, is flaming whereby the bully publicly initiates an online fighting by exchanging emails/posts containing insults and indignity, and sometimes sending abusing images to the victims [11]. If the exchanged messages are discriminative and used to exclude the victim from the community or group, this type is called exclusion, which is considered more dangerous as it leads the victim to be alone and isolated [34].

Above and beyond, posting a private content about the victim publicly is another cyberbully form which has two types: outing and denigration. Outing involves the bully publicly sharing a private information, including photos or videos, of victim with the intention of harm, while denigration focuses on sending harmful or hurtful content about the victim publicly [12]. Overall, all these types of cyberbully are harmful to the victims, who react differently based on the cyberbully situation, as will be explained in the next section.

2.2. Cyberbullying: Reactions and Effects

Cyberbullying represents a real danger to the victims, who react differently based on their culture, growth and/or societal perception. There are four different possible reaction categories against cyberbullying based on the victim personality. In reality, the growth and gender may form the reaction against the bully. One of the positive reactions, which shows the maturity of victim [34] is the active reaction whereby the victim tends to share the case with responsible adults like parents and teachers seeking for assistance and protection. This form is one of the directions guided and recommended by the intensive awareness programs [20]. On the other side, some victims has the ability to face the issue and take a kind of revenge against the bully as a self-defense, this reaction called violent behavior [13], which is a negative reaction where the victim will be a potential bully who will insult and abuse others [33]. The rest of victims are trying to follow the evasion or passive reactions. In evasion reaction, the victim tries to tolerate the cyberbully with no actions [13]. The passive reaction of victim means trying to avoid the bully by deleting the messages or links sent by bully. The last two reactions may lead the victim to further psychological complications [33,34]. On the whole, the victim will live in an inner struggle situation, which might lead to psychological conditions such as the psychological inhibition. According to [13], the cyberbullying victims tend to maintain their privacy by tolerating cyberbullying or avoiding the bully without making a noise. Fewer cases are reported to be shared with parents and teachers. In a conservative society such as Saudi Arabia, the victims feel ashamed of disclosing the bully cases to avoid the scolding by parents/elder relative. In fact, one of the fatal mistakes done by parents is ignoring or scolding the victim to make himself/herself independent. In addition, some parents feel disappointed that their kid has no capability to solve the issues by himself/herself [14]. All these reactions from both victims and parents lead to unpredictable escalation of cyberbully.

With the amassed escalation of this social attack, cyberbullying has numerous effect on the victims. With no way to avoid, cyberbullying penetrates victim’s life on a daily basis as the victims are staying on the Internet. Apart from technology addiction, cyberbullying has become one of the societal issues that affects the victims with severe impacts including social isolation, low self-esteem and poor self-image [15]. These issues are not only individual issues, but also threaten the societal relations. Moreover, cyberbullying is one of the main reasons of anxiety, paranoia and depression [16]. These psychological conditions may lead to addiction and alcohol/drugs consumption. As a final stage of suffering, this might lead to suicide commitment [17]. The cyberbullying victim may feel unsafe wherever he/she goes, which leads him/her to loneliness and insomnia. Consequently, the victims may have physical health conditions such as abdominal pain and headache. According to a statistical study on a random sample of school students [18], 42.5% of cyberbullying victims feel irritated, ∼40% of them feel annoyed and about 27% of the victims feel sad.

Interestingly, the impact of cyberbullying is not restricted to the victim only, but it is also propagated to the bully. According to some psychological studies, represented by [6], most of the bullies have psychological issues such as frustration, jealous of the victim and/or arrogance. These conditions motivate the bully to perform cyberbully attack against others to feel satisfied. In some cases, the bully was a victim and he/she tries to play the role of bully [19].

Despite the negative consequences of cyberbully, there is no way to prevent cyberbullying. Most parents and teachers rely on awareness of children on the causes and impacts of cyberbullying. As well, some parents think that peer-mentor is an effective way to prevent cyberbullying, particularly in the teenage years where the peers have massive impact than family and school [20].

2.3. Anti-Cyberbully Rules and Regulations

Social media platforms established mainly on the principle of ‘Freedom of speech’, which enables the individuals to express their opinions freely without a fear. However, this freedom is restricted with some prohibitions. For instance, the social media users should not include any sensitive contents in their posts. These sensitive contents include sexuality, politics or abusive behavior. In fact, Twitter, as a dominant social media platform, follows this principle by letting the users interact freely; however, this freedom is controlled by monitoring the users’ behavior and contents [21]. This practice belongs to the restricted freedom of speech. Out of the restrictions on the freedom of speech, Twitter forbids those contents containing abusive behavior to maintain a healthy and safe social communication environment. Merriam-Webster [3] defines abusive behavior as using harsh or insulting language against others.

Therefore, any post on social media containing such harsh or insulting language would be an abusive behavior. Abusive behavior on social media includes harassment, sharing sensitive content, self-harm, violent threats and so on. In the abusive behavior cases with low severity such as hateful conduct, Twitter requests the offender to remove the post, preventing to post any new contents. In the medium severity cases such as harassment, Twitter blocks the account with inability to log in to the account. In the high severity cases such as violent threat, Twitter deletes the account permanently [21]. Despite immediate actions by Twitter against the offenders to reduce the impact of abusive behavior, these rules are insufficient and subjective. For example, if tens of users report your account as harmful account, your account will be locked, even if you do not do any abusive behavior.

From the legal perspective, there are a number of global initiatives to prevent cyberbully and enhance the safety of Internet users, especially the children. For instance, university of Tuku, Finland has established anti-cyberbully program called Kiva (http://www.kivaprogram.net), anti-harassment campaign in France (https://www.nonauharcelement.education.gouv.fr/) and anti-cyberbully initiative by Belgian government (https://www.veiligonline.be/cyberpesten).

In Saudi Arabia, cyberbullying is classified legally a cybercrime, and the bully is subject to legal prosecution under The Saudi Anti-Cyber Crime Law. The stated punishments in this law range between 1 and 5 year imprisonment, as well as, payment of SAR500000 to SAR3000000 based on the severity of crime. Despite tough sanctions in Saudi Arabia against cybercrime including cyberbullying, cyberbullying escalates tremendously, particularly among school and university students. This might be due to the conservative nature of Saudi society, as explained earlier. There is no way to confirm that enforcement of laws helps in preventing cyberbully issues; this will require a lot of monitoring to conduct the analysis [22]. Thus, the need for accurate and automated cyberbullying detection techniques without any bias from social media users is still needed.

2.4. Discussion on the related works

Cyberbully is present in the research arenas, and the researchers work intensively to detect the cyberbully in order to find a way to control or reduce the cyberbully in social media. One direction in this field is to detect the user's intention to post offensive contents by analyzing the used offensive language based on different features like the structure and the unique content of cyberbully, as well as, the users' writing style. Chen et al. [23] use Lexical Syntactic Feature (LSF) approach, which showed the ability to detect the offensive users with precision of 77.9%, and recall of 77.8%. The authors in [24] proposed a platform from pluggable and reusable components to trace the harmful content on social media. Their platform analyzes the content based on the type (video, audio, text and/or images), and once the inappropriate content is detected, a report will be sent to the social media moderator to investigate and flag the profile as offender. There is neither sentiment analysis used nor results shown in this research. Li et al. [25] proposed a Twitter based Event Detection and Analysis Systems (TEDAS) whereby the events will be detected based on the rich information retrieved from Twitter. In this system, the tweets are retrieved via Twitter API, and are classified based on a predefined Crime and Disaster related events such as care accidents, flood, building collapse and so on. The system stores the events using meta-information extractor along with their geolocation and temporal label into database. The user can search for any event through an interface. The system can be used for the events related to cyberbully and other societal issues with certain enhancements. Hua et al. [26] proposed a similar system called Semi-supervised Targeted Event Detection (STED) with adding extra search features to let the users find the target event with a high accuracy.

Utilizing machine learning in cyberbully detection has been presented by different authors [11,35–39]. The authors in [37] utilized machine learning to detect the cyberbully in Turkey. The authors build their own small dataset. Besides, the authors in [11] built their own dataset to detect cyberbully in Dutch language. The authors in [35] detect the cyberbully from YouTube comments. According to [33], both studies [11,35,37] have limitation in the features used and they are unable to recognize the bully and victim. The authors in [36,38] incorporated the cyberbully detection with text features.

All the aforementioned works focus on text analysis without considering the sentiment analysis to predict the bully intention. Thus, based on the aforementioned works, the author can claim that it is possible to detect cyberbullying from social media. In this proposed research, the author will use sentiment analysis to detect cyberbullying. It is worth mentioning that in the current decade, Twitter sentiment analysis has attracted many researchers and practitioners due to its promising results. Usually, sentiment analysis is applied in the business domain to understand customer behavior or feedback [27]. Numerous research and practical papers have proposed solutions based sentiment analysis for different business domains and recently sentiment analysis has been used to provide solutions in the social sciences as well. In this study, sentiment analysis is used as a research tool for the field of cyberbullying.

3. PROPOSED CYBERBULLYING DETECTION FRAMEWORK

Detecting cyberbullying on social media through the analysis of users' behaviors are both theoretical and practical challenges. From the theoretical viewpoint, the researchers rely on fact-finding techniques such as questionnaires and interviews to stem the perception of participants and to analyze the impact of cyberbullying on a certain discipline, e.g. educational progress.

TABLE 1. A Sample of Keywords used in Cyberbully Context.

Curse	No one ever likes you	He is the ugliest of all, all of his fans are stupid
Stupid	You're ugly!	Get lost
Beat you	Chink	Mozzah
Fat/belly/big stomach	Ask you go die	I will send you to hell
Pimple faced	You are idiot	Go to die
Fat ass/fatso/slowpoke etc.	'Your mom is a whore, slut, etc. '	Stupid, idiot
Ugly	Bitch	Fuck off
'You're the toxic of this world.'	Noob	Ma Had Hawlak
Rude	Insulting	Ethlef
Insult	Humiliation	Kol Zegg
Pull of shit, shit	You should go kill yourself	You fat shit
Stupid	Your face so sucks	It depends on the individual.
Threatening words	Idiot	Hot ass

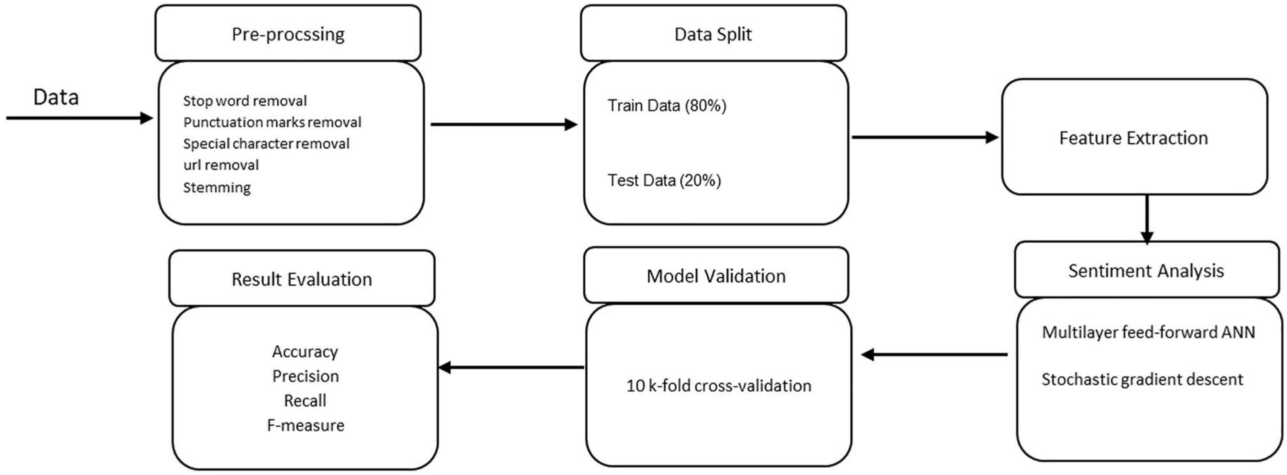

FIGURE 1. The proposed Cyberbully Detection Framework.

In this research, the fact-finding techniques were used to extract the keywords of cyberbullying that align with Middle East context. The keywords were extracted from both experts and ordinary participants to maintain the diversity of coverage. Table 1 shows sample keywords collected from the fact-finding techniques.

After collecting the keywords, these keywords were used as an indicator of Cyberbully. In the next step, each tweet will be annotated manually as 'positive' or 'negative' based on its inclusion of Cyberbully keywords. Therefore, the methodology shows the unification of polarity. The proposed system is described in the following section.

The proposed system has been built using Visual Studio 2017 due to the availability of modules, ease to connect with Twitter API and flexibility of integration. Tweetinvi is a great extension provided by Visual Studio 2017 that provides the reliable and easy interaction with Twitter through the Twitter official API. The proposed Cyberbullying detection system has four phases: the preprocessing phase, feature extraction phase, classification phase and evaluation phase. Figure 1 depicts the phases of proposed system.

3.1. Data Streaming from Twitter

Using the fact-finding techniques (questionnaire for ordinary people & interview with experts), a group of keywords have been identified to represent the cyberbully in Middle East context. The proposed model is expected to detect the cyberbully based on these keywords. Afterward, the authors started working on streaming the tweets from Twitter. Twwetinvi has been used to connect with Twitter streaming API in order to retrieve the global tweets streams according to our constraints (location, keyword filtering and language filtering). The sample code of Twwetinvi can be found in github [https://github.com/linvi/tweetinvi]. Therefore, the author conducted a search process to limit the tweets within boundaries of Saudi Arabia according to geo-coordinates. The resulting tweets were generated as records and each record consisted of {Tweet ID, Tweet Text and retweet count}. As is well known, the text in Twitter includes many special use symbols and phrases such as http, emoji faces, hash tag symbols, mention symbols and retweet, etc. Thus, the preprocessing is required to eliminate such noise.

3.2. Data Preprocessing and Cleaning

To make the data ready for classification, the preprocessing is a must to remove and clean unwanted noise in text detection. The data preprocessing include [39] cleaning of texts (e.g. removal of stop words and punctuation marks) and spam content removal. In the proposed model, it has been applied to remove and clean unwanted noise in text detection. Such text noise includes stop words, special characters and repeated words. Then, the stemming for the remaining words to their original roots has been applied as a result of this preprocessing, and the dataset containing clean tweets is produced for the proposed model to be run and predicted. Besides, the author removed retweets (−rt) and links (−http) to reduce noise and duplications.

The output of this step was a dataset in which each tweet was represented as a record. The dataset containing clean tweets produced for the proposed model to be run and tested was stored in MongoDB server, which runs locally. For this purpose, Happier Fun Tokenizer [https://github.com/dlatk/happierfuntokenizing] was used. This tokenizer will produce clean tweets with no symbols, URLs, Phone numbers or repetition.

3.3. The proposed Sentiment Model

To build the sentiment model, the assumption that the Cyberbully is subjective and contextualized was employed. For example the sentence ‘You are pretty!!!!! hahahaha’, this sentence is normal sentence if there is no exclamation or laughing at the end. Thus, the manual annotation is used based on the experts and the potential victims’ opinions. In the proposed systems, the fact-finding (or crowdsourcing) was used to build keywords pre-list. An interface was designed to allow the users to manually annotate the tweets sentiment for training set as positive or negative.

The proposed classification model was developed based on deep learning [28] as a multilayer feed-forward artificial neural network that was trained with stochastic gradient descent using a backpropagation algorithm. The aim of this step was to build a model that could predict whether the new tweets were positive (contains cyberbullying) or negative (does not contain cyberbullying). The result was cumulative and based on the majority. As our aim was to identify cyberbullying, the output result reflected the cyberbullying situation.

Furthermore, TF-IDF feature Extraction was used. TF-IDF is a combination of TF and IDF (term frequency-inverse document frequency), and this algorithm is based on word statistics for text feature extraction. This model considers only the expressions of words that are the same in all texts. Therefore, TF-IDF is one of the most commonly used feature extraction techniques in text detection. Further discussion on TF-IDF can be found in [39].

Next, the developed model was validated by employing 10 k-fold cross-validation [29] to test the prediction accuracy of the developed model on the dataset. Cross-validation split the data into k subprocesses. Each time one subprocess was used as a validation set; the remaining k-1 subprocesses were used as the training set. The results obtained from the k experiments were averaged to get a single validation value. The output of this step was a validated classification model.

4. RESULTS AND DISCUSSION

The main goal of this paper is to investigate the feasibility of automatically detect the Cyberbully from the users posts in Tweeter using sentiment analysis model. In this section, the recorded results are dedicated to evaluate the ability of proposed model to detect the Cyberbully tweets. Therefore, the following performance metrics were used: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision, Recall and Accuracy, as explained in the following lines.

- True Positive (TP): the tweet is correctly detected as Cyberbully
- False Positive (FP): the tweet is incorrectly detected as Cyberbully.
- True Negative (TN): the tweet is correctly detected as NOT Cyberbully
- False Negative (TN): the tweet is incorrectly detected as NOT Cyberbully
- Precision calculates the ratio of relevant tweets in the middle of the true positive (TP) and false positive (FP) tweets belong to a particular category

$$\text{Precision} = \text{TP}/\ (\text{TP} + \text{FP}) \tag{1}$$

- Recall calculates the ratio of retrieved relevant tweets over the total number of relevant tweets

$$\text{Recall} = \text{TP}/\ (\text{TP} + \text{FN}) \tag{2}$$

- Accuracy calculates the ratio of the true detected cases to the overall cases

TABLE 2. The Performance Evaluation of Proposed Model.

	TP	TN	FP	FN	Recall	Precision	F-score	Accuracy
The proposed model	228	17	43	12	0.95	84.1	0.89	0.816
CISD [32]	210	25	40	25	0.89	0.84	0.86	0.78

$$\text{Accuracy} = \text{TP} + \text{TN}/\ (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \qquad (3)$$

Table 2 summarizes the results out of 300 tweets. From the table, 245 out of 300 test data had successfully predicted the same intent as the labeled intent in the sample file.

The proposed model was compared with a similar work proposed by [32]. The experiment has been conducted on the two models to evaluate the superiority of the proposed model. The results are summarized in Table 2. Out of 300 tweets, 76% of tweets were detected as cyberbully (true positive) compared with 14% of tweets wrongly detected (False positive). Besides, the accuracy of the proposed model is 81%, which is higher than the accuracy of SCID model proposed by [32], which is 78%. Both models were tested on the same dataset. The interpretation of high accuracy of the proposed model is due to the model validation component using 10 k-fold cross-validation that validates the detected cyberbully. Such validation eliminates the wrongly detected cyberbully. Table 2 shows the values are close to each other with a slight difference depicting the superiority of our proposed model.

5. CONCLUSION

Twitter is the prominent social media platform in Saudi Arabia with huge community who interacts, communicate and share knowledge. As a side effect of using social media, abusive behavior emerged, particularly cyberbullying. In Saudi Arabia, cyberbullying is widespread, especially in university and school students. Despite tough sanctions of cyberbully, it is still a problem for victims and society. This paper introduces a cyberbullying detection methodology, which extracts the tweets from Twitter and classifies the tweets, based on extracted keywords from both experts and potential victims, using deep-learning classification algorithm. The main aim of this paper is to use sentiment analysis technique that can detect the cyberbully intention in the tweet stream. The results show the ability of proposed methodology to detect cyberbullying effectively with accuracy 81.6%. Substantial future work is highly needed to identify the keywords suitable for Middle East context effectively without users' bias. **Therefore, to make this work more realistic, the author continues working on the cyberbully detection in the Arabic context by building Arabic dataset, which will be extracted from the tweets reflecting the intention of Tweeter users' majority in Saudi Arabia rather than focusing on only university and schools' students. Besides, the future work includes extracting the most influential features through investigating the feature extraction that can produce high accuracy in cyberbully detection from Arabic tweets. Another direction in future work is evaluating different classification models to use the most accurate classifier to enhance the model accuracy. By integrating the feature extraction and the accurate classifier on Arabic dataset, the author expects the cyberbully detection will be feasible. Moreover, another future direction is to work on real time detection of cyberbully using automated machine learning (TPOT), which automates the classification process whereby multiple classifiers are optimized and utilized to enhance the accuracy**.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

SUPPLEMENTARY MATERIAL

Supplementary material is available at www.comjnl.oxfordjournals.org.

ACKNOWLEDGEMENT

This work was supported by Prince Sultan University, Riyadh, KSA. The author would like to thank Prince Sultan University, Riyadh, KSA, for partially supporting this work.

DISCLOSURE

The authors declare that there is no conflict of interest.

REFERENCES

[1] Edosomwan, S., Prakasan, S.K., Kouame, D., Watson, J. and Seymour, T. (2011) The history of social media and its impact on business. *Journal of Applied Management and entrepreneurship*, 16, 79–91.

[2] Global media insights (2019). *Saudi Arabia Social Media Statistics 2019*, Dubai, UAE. Available at: https://www.globalmediainsight.com/blog/saudi-arabia-social-media-statistics/ [Accessed 01 June, 2020]

[3] Merriam-Webster (2019). *Social Media* | Definition of Social Media by Merriam-Webster. Available at: https://www.merriam-webster.com/dictionary/social%20media [Accessed 01 June, 2020].

[4] Bryce, J. and Fraser, J. (2013) 'It's common sense that it's wrong': Young people's perceptions and experiences of cyberbullying. *Cyberpsychol. Behav. Soc. Netw.*, 16, 783–787.

[5] Smith, P.K. (2011) Cyberbullying and Cyber aggression. In Jimerson, S.R., Nickerson, A.B., Mayer, M.J., Furlong, M.J. (eds) *Handbook of School Violence and School Safety: International Research and Practice*. Routledge, New York.

[6] GÖrzig, A. and Ólafsson, K. (2013) What makes a bully a cyberbully? Unravelling the characteristics of cyberbullies across twenty-five European countries. *J. Child. Media*, 7, 9–27.

[7] Calvete, E., Orue, I., Estévez, A., Villardón, L. and Padilla, P. (2010) Cyberbullying in adolescents: Modalities and aggressors' profile. *Comput. Hum. Behav.*, 26, 1128–1135.

[8] Hoff, D.L. and Mitchell, S.N. (2009) Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration*, 47, 652–665.

[9] Jones, S.E., Manstead, A.S.R. and Livingstone, A.G. (2011) Ganging up or sticking together? Group processes and children's responses to text-message bullying. *Br. J. Psychol.*, 102, 71–96.

[10] Notar, C.E., Padgett, S. and Roden, J. (2013) Cyberbullying: A review of the literature. *Univ. J. Educ. Res.*, 1, 1–9.

[11] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. and Hoste, V. (2018) Automatic detection of cyberbullying in social media text. *PLoS One*, 13, e0203794.

[12] Li, Q. (2010) Cyberbullying in high schools: A study of students' behaviors and beliefs about this new phenomenon. *J. Aggress. Maltreat. Trauma*, 19, 372–392.

[13] Wong, D.S., Chan, H.C.O. and Cheng, C.H. (2014) Cyberbullying perpetration and victimization among adolescents in Hong Kong. *Children and youth services review*, 36, 133–140.

[14] DeHue, F., Bolman, C. and VÖllink, T. (2008) Cyberbullying: Youngsters' experiences and parental perception. *Cyberpsychol. Behav.*, 11, 217–223.

[15] Cowie, H. (2013) Cyberbullying and its impact on young people's emotional health and well-being. *The Psychiatrist*, 37, 167–170.

[16] Schenk, A.M. and Fremouw, W.J. (2012) Prevalence, psychological impact, and coping of cyberbully victims among college students. *Journal of school violence*, 11, 21–37.

[17] Carma, H. (2016) *Bullied teen kills herself in front of her family*. CNN.com, United States of America.

[18] Vandebosch, H. and Van Cleemput, K. (2009) Cyberbullying among youngsters: Profiles of bullies and victims. *New Media Soc.*, 11, 1349–1371.

[19] Miller, A. (2017) *Couple arrested for bullying Texas girl who shot herself* | Daily Mail Online. Daily Mail UK, United Kingdom.

[20] Steinmetz, J.M. (2013) *Cyberbullying and the digital divide: Student and teacher perceptions and reactions* Doctoral dissertation. The Ohio State University, Columbus, United States.

[21] Twitter (2019). *The Twitter Rules* | Twitter Help Center. [Online] Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules [Accessed 20 May, 2019].

[22] Coburn, P.I., Connolly, D.A. and Roesch, R. (2015) Cyberbullying: Is federal criminal legislation the solution? *Can. J. Criminol. Crim. Justice*, 57, 566–579.

[23] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing (pp. 71–80)*. IEEE, Amsterdam, Netherlands.

[24] Vanhove, T., Leroux, P., Wauters, T. and Turck, F.D. (2013) Towards the design of a platform for abuse detection in OSNs using multimedial data analysis. *IM*, 2013, 1195–1198.

[25] Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012). Tedas: A twitter-based event detection and analysis system. In *2012 IEEE 28th International Conference on Data Engineering (pp. 1273–1276)*. IEEE, Arlington, Virginia USA.

[26] Hua, T., Chen, F., Zhao, L., Lu, C. T., & Ramakrishnan, N. (2013). STED: semi-supervised targeted-interest event detection in twitter. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1466–1469)*. ACM, Chicago, USA.

[27] Neuendorf, K.A. (2016) *The content analysis guidebook*. Sage, USA.

[28] Bengio, Y., Courville, A. and Vincent, P. (2013) Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35, 1798–1828.

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

[30] Simon K. (2020), *DIGITAL 2020*: SAUDI ARABIA. https://datareportal.com/reports/digital-2020-saudi-arabia [Accessed on 27 May, 2020].

[31] Alotaibi, N.B. (2019) Cyber bullying and the expected consequences on the students' academic achievement. *IEEE Access*, 7, 153417–153431.

[32] Dani, H., Li, J., & Liu, H. (2017). Sentiment informed cyberbullying detection in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 52–67)*. Springer, Cham.

[33] Talpur, K.R., Yuhaniz, S.S. and Amir, N.N.B. (2020) Cyberbullying detection: Current trends and future directions. *J. Theor. Appl. Inf. Technol.*, 98, 3197–3208.

[34] Reyns, B.W. and Fissel, E.R. (2020) Cyberstalking. *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, pp. 1283–1306. Springer Nature, Switzerland.

[35] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *In Proceedings of the Social Mobile Web*. The AAAI Press, Spain.

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

[37] Özel, S. A., Saraç, E., Akdemir, S., & Aksu, H. (2017, October). Detection of cyberbullying on social media messages in Turkish. In *2017 International Conference on Computer Science and Engineering (UBMK) (pp. 366–370)*. IEEE, Antalya, Turkey.

[38] Huang, Q., Singh, V. K., & Atrey, P. K. (2014). Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia (pp. 3–6)*. ACM, Orlando, Florida, USA..

[39] Muneer, A. and Fati, S.M. (2020) A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12, 187.