Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Automatic detection of cyberbullying and threatening in Saudi tweets using machine learning

Deema Alghamdi, Rahaf Al-Motery, Reem Alma'abdi, Ohoud Alzamzami *, Amal Babour

Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 26 January 2021
Received in revised form
25 May 2021
Accepted 28 June 2021

Keywords:
Artificial intelligence
Arabic language
Cyberbullying
Text classification
Machine learning

ABSTRACT

Social media has become a major factor in people's lives, which affects their communication and psychological state. The widespread use of social media has formed new types of violence, such as cyberbullying. Manual detection and reporting of violent texts in social media applications are challenging due to the increasing number of social media users and the huge amounts of generated data. Automatic detection of violent texts is language-dependent, and it requires an efficient detection approach, which considers the unique features and structures of a specific language or dialect. Only a few studies have focused on the automatic detection and classification of violent texts in the Arabic Language. This paper aims to build a two-level classifier model for classifying Arabic violent texts. The first level classifies text into violent and non-violent. The second level classifies violent text into either cyberbullying or threatening. The dataset used to build the classifier models is collected from Twitter, using specific keywords and trending hashtags in Saudi Arabia. Supervised machine learning is used to build two classifier models, using two different algorithms, which are Support Vector Machine (SVM), and Naive Bayes (NB). Both models are trained in different experimental settings of varying the feature extraction method and whether stop-word removal is applied or not. The performances of the proposed SVM-based and NB-based models have been compared. The SVM-based model outperforms the NB-based model with F1 scores of 76.06%, and 89.18%, and accuracy scores of 73.35% and 87.79% for the first and second levels of classification, respectively.

© 2021 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In recent years, social media has been widely used around the world. The role of using social media is to allow people to communicate, exchange messages, share knowledge, and interact with each other. Social media use has become an increasingly popular component of everyday activities. Due to that, a huge amount of data on social media websites and microblogs, such as Twitter and Facebook, are being added every day (Altaher, 2017). A study has shown that Saudi Arabia has the highest annual growth rate of social media users around the world (Alruily, 2020). With Twitter users posting about 500 million tweets per day, over 30% of these tweets are from Saudi Arabia (Alruily, 2020).

With this growth of social media websites and the increasing number of users, the forms of abuse and violence have evolved from the real world to the virtual world. Although social media have helped to connect people around the world, some people abuse this technology by using violent texts to verbally attack other users in many ways, such as bullying, insulting, swearing, and extortion. Violent text is abuse that takes place over digital devices like cell phones, computers, and tablets (Haidar et al., 2016). Social media violence, such as cyberbullying, can have a negative effect on people's psychological and mental health that could even be worse than physical violence, especially for teenagers and young people. Cyberbullying can spread at a wider scale than real-world bullying, in addition, violent text posted on social media is left there forever, which could have a long-term effect on people unless these posts are reported and removed. Many of the individuals who are affected by social media violence do not report such incidents for several reasons, which include fearing that things will get worse or being under threat by the bully who prevents them

* Corresponding Author.
Email Address: ualzamzami@kau.edu.sa (O. Alzamzami)
https://doi.org/10.21833/ijaas.2021.10.003
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-6555-8166
2313-626X/© 2021 The Authors. Published by IASE.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

from reporting these incidents. Moreover, due to a large number of users and the huge amounts of social media data, which are generated on a daily basis, it is difficult to detect, track, and stop such kinds of attacks manually.

Thus, the aim of this paper is to automatically detect violent tweets on Twitter. Compared to existing papers which have been proposed in the literature for violent Arabic text detection and classification (Ismail et al., 2018; Duwairi and Qarqaz, 2014; El-Naggar et al., 2017; Biltawi et al., 2017; Mouheb et al., 2018; Haidar et al., 2017), this paper is focusing on the Saudi dialect. The main contribution of this paper is to build a two-level classification model for violent text as follows:

- First level of classification classifies the tweets into either violent or non-violent text.
- Second level of classification classifies the violent text into either cyberbullying or threatening.

The rest of this paper is organized as follows: Section 2presents the related work on text classification, Section 3 describes the proposed methodology, Section 4 presents the experiments, results, and discussion. Finally, Section 5 concluded this paper.

2. Related works

Several works on detecting hate speech and offensive language have been done in many languages such as English (Gambäck and Sikdar, 2017; Mahmud et al., 2008; Spertus, 1997), German (Wiegand et al., 2018; Schneider et al., 2018; Ross et al., 2017), Hindi (Kumar et al., 2018; Modha et al., 2018), Mexican Spanish (Díaz-Torres et al., 2020), Dutch (Van Hee et al., 2015), and Arabic c (Ismail et al., 2018; Duwairi and Qarqaz, 2014; El-Naggar et al., 2017; Biltawi et al., 2017; Mouheb et al., 2018; Haidar et al., 2017). A review of techniques used in Arabic language cyberbullying detection, including natural language processing, and machine learning have been presented in Haidar et al. (2016). Using a set of 175 million Arabic tweets collected during March 2014, a list of obscene words was extracted to be used in identifying offensive text content (Mubarak et al., 2017). These words were used to build a corpus of 660 thousand Arabic offensive tweets, which were collected between April 15, 2019, and May 6, 2019 (Mubarak et al., 2020). About 10 thousand of these tweets were annotated manually by experienced annotators (Mubarak et al., 2020). The authors provided the annotators with a set of guidelines to help them in labeling the tweets as either offensive or clean where offensive tweets include vulgar and hate speech (Mubarak et al., 2020). Likewise, a dataset of 15,050 Arabic comments on celebrities in the Arab world was collected from YouTube videos in July 2017 (Alakrot et al., 2018). The comments were annotated as either offensive or inoffensive by three annotators from different Arab countries (Alakrot et al., 2018). Further, a Support Vector Machine (SVM) classifier was used on the prepossessed dataset with/without stemming and with/without normalization and the results showed that data preprocessed dataset with stemming can enhance the detection of offensive comments (Alakrot et al., 2018).

Detecting and classifying cyberbullying Arabic tweets in real-time based on their strength was proposed in Mouheb et al. (2019). The authors created a list of offensive words with three different classes, which are mild, medium, and strong. If a comment contains any word from the offensive words list, it is classified as cyberbullying. In addition, detected cyberbullying tweets were classified based on their strengths by assigning a weight function for each comment. The weight function considers the number of bullying words in the comment and the weight of each word. A dataset of 100,327 tweets and comments were collected from Microsoft Flow and YouTube and classified as either cyberbullying or not based on lexicon using Pointwise Mutual Information (PMI), Chi-square, and Entropy approaches (AlHarbi et al., 2019).

Multiple classifiers were applied to Arabic text to detect offensive language. A single learner machine learning (SVM, logistic regression, and decision tree) and ensemble machine learning (bagging, Adaboost, and random forest) were applied on Arabic offensive tweets collected in Al-Khalifa et al. (2020) for the purpose of detecting offensive language in Arabic text (Husain, 2020). The results showed that ensemble machine learning achieved better results than single learner machine learning and bagging ensemble machine learning classifiers was the best in detecting offensive language. A comparison of four neural network classifiers, which are Convolutional Neural Network (CNN), Bidirectional Long Short Term Memory (Bi-LSTM), attention Bi LSTM, and a combined CNN-LSTM on a was done in Mohaouchane et al. (2019). The data set used was created in Alakrot et al. (2018) and the results showed that the combined CNN-LSTM achieved the best recall and the CNN achieved the best accuracy and precision among the classifiers for detecting offensive on Arabic social media (Mohaouchane et al., 2019).

3. Methodology

The proposed methodology consists of five main steps, as illustrated in Fig. 1. The first step involves collecting the required data to build the classifier model. Then, the data is preprocessed and annotated to train the model. After that, the classification is done in two levels, where the first level of classification classifies the tweets into either violent text or non-violent. Further, the violent text is classified using the second classifier into cyberbullying or threatening.

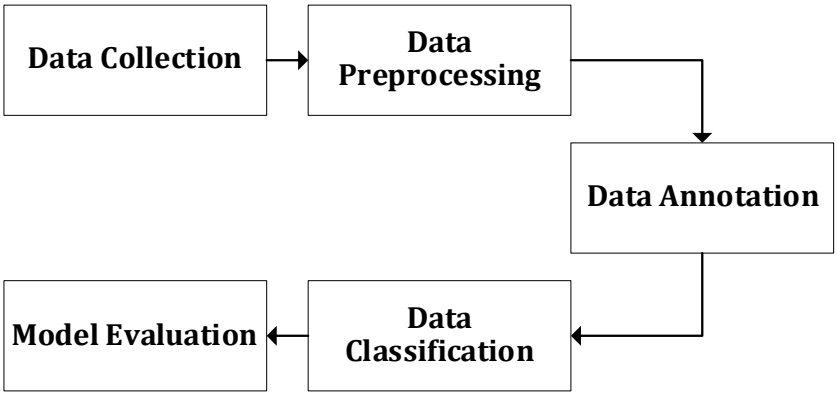

Fig. 1: Research methodology

3.1. Data collection

Twitter is a social website where people write their opinions and thoughts about different topics, which makes Twitter rich with text data. With the help of Twitter API and Tweepy which is a library available in the python language, the total number of tweets that have been collected is 3700. Those tweets are collected using 50 keywords and common hashtags in Saudi Arabia, in addition to using different filter settings to ensure that the collected dataset contains a sufficient number of violent texts from both categories, cyberbullying and threatening. Table 1 and Table 2 show samples of the cyberbullying and threatening keywords, respectively, which are used to collect the dataset from Twitter. Moreover, there are some keywords that did not indicate cyberbullying by itself such as "mryḍ," which means being sick. When this word is used alone, it retrieves a normal tweet such as "āllhm āšfy kl mryḍ," which is a prayer for a sick person to recover from illness. However, when adding some prefix to it such as adding “yā” to “mryḍ” it becomes “yāmryḍ,” which is an insulting text that means "you are sick." Using the resulting term “yāmryḍ” would retrieve cyberbullying tweets. In addition, in the Arabic language, a word may be written in different forms such as "dbh’" and "dbh," which means "you are fat like a bear." The different forms of the same words are added to the list.

Table 1: Searching keywords for threats

Root	Meaning	Examples of root forms		
fḍḥ	shame	ḥfḍḥk	bfḍḥk	fḍyḥ'
l‘n	curse	āl‘nk	bl‘nk	tnl‘nyn
ğld	lash	bğldk	ḥtnğld	btnğldyn
šhwr	expose	bāšhrk	ḥāšhrk	bšhrk
qtl	kill	bqtlk	ḥāqtlk	btnqtl
ṭ‘n	stab	-aṭ‘nk	bāṭ‘nk	ḥāṭ‘nk
šwh	deform	bšwhk	ḥāšwhk	ḥšwhk

Table 2: Searching keywords for cyberbullying

Root	Meaning	Examples of root forms		
‘abada	niger	‘bwd	yā‘bd	yāl‘bd
db	fat	yādb	dbh’	dbh
mrḍ	sick	mryḍ	yāmryḍ	yāmryḍh’
qrf	disgusting	mqrf’	yāmqrf	mqrf

Also, common violent phrases in Saudi Arabian dialect were added to both lists. For example, "ābn āmk wryny wğhk," which "means show me your face if you dare" and has been added to the threatening list. Table 3 shows some examples of these sentences.

Table 3: Examples of threatening sentences

Threatening Sentences	Meaning
ābn āmk t‘āl	come if you dare
ābn āmk wryny wğhk	show me your face if you dare
wāllh lād‘s ‘lY wğhk	I swear I will step on your mouth

The retrieved tweets are saved in two excel files. Some of the tweets which include non-Saudi dialect, advertisement, and non-text contents have been manually removed from the dataset. Thus, after the cleaning process, the resulting dataset contains 2000 tweets which include both violent text and non-violent text.

3.2. Data annotation

Using guidelines from two experts in psychology and a handbook from the "Be Free" program of the Bahrain women’s association, named “Say no to cyberbullying,” the tweets are annotated as either normal, cyberbullying, or threatening. To annotate the tweets, a copy of the tweets associated with the guidelines was sent to two annotators. A third annotator is involved only when there is a disagreement between the two participants as the final label for each tweet. After the annotation process is done, the agreement between the annotators has been calculated to ensure the reliability and quality of the annotation process using Cohen’s Kappa agreement, which considers the fact the annotators may disagree or agree by chance (Vieira et al., 2010; Al-Kabi et al., 2016). A substantial agreement of 80% has been found between the annotators. Table 4 shows examples of annotated tweets based on the guidelines. Table 5 summarizes the number of annotated tweets for each class, non-violent, cyberbullying, and threatening.

3.3. Data pre-processing

The pre-processing phase involves four main tasks, which are tokenization, noise removal, normalization, and stop-word removal. Fig. 2 shows the workflow of the pre-processing step.

Table 4: Examples of the annotated dataset

Arabic Tweet Text and Meaning	Classification
dbh ūtdḫl al-āṣnṣīr ?????? anqlʿī amšī fī al-drǧ īā ʿāhh ūḫffī sʿrāt ūbqī lnā mkān al-lh lāīūfqk.	Violent text Cyberbullying
You are a fat girl who should not use the elevator Get out! Use the stairs and burn some calories leave some space for us in the elevator.	
ūrbī lʾaqtlk īālʿbdh lā tṣīrī krīhť trā mā ʿndī mškh ablkk.	Violent text Threat
I swear I will kill you, niger don't be a skunk, or else I will block you.	
ʿndī suʾāl lnās al-mṯqfť wālmḥbť llqrāʾť anā ǧāls afkr anī aṯqf mn nfsī wāstfīd mn ūqtī wudī aqrāʾ lī ktāb ūš tnṣḥūnī ṭbʿāǒrāḥ īkūn aūl ktāb.	Normal (Non-Violent text)
I have a question for people who love to read I am thinking that I should educate myself and use my time efficiently in reading. Would you give some advice on the first book I should read?	

Table 5: Number of annotated tweets for each class

Class		Number of Tweets	
Violent	Cyberbullying	583	1010
	Threatening	427	
Not Violent		990	

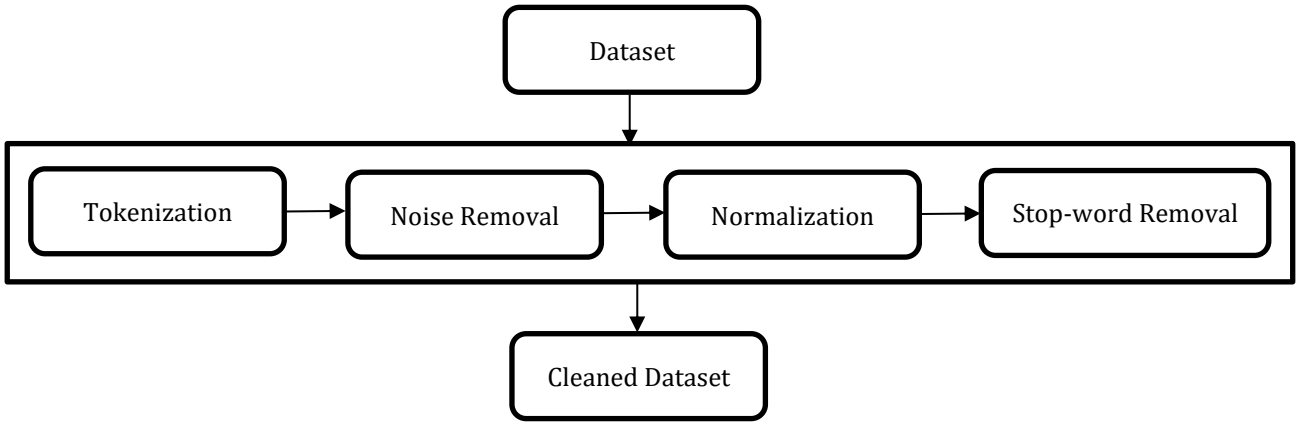

Fig. 2: Summary of data pre-processing

- Tokenization: The first step of pre-processing is tokenization. This tokenization is the process of breaking up the tweets into separate words based on the space, comma, semicolon, colon, and dot. The main benefit is to deal with each word separately which makes the features extraction process for the next cycle easier. For example, after applying tokenization to the following sentence:

"trāk qzm mṯlhā lā tswy fyhā hhhhhhhhhhhhhhhhhhhhhhhhhhhhh"

which means "you are a dwarf. Do not overestimate yourself; LOL!," the generated tokens are as follows:

"trāk," "qzm," "mṯlhā," "lā," "tswy," "fyhā," "hhhhhhhhhhhhhhhhhhhhhhhhhhhhh"

tokenization was done using the Natural Language Toolkit (NLTK) library in Python.

- Noise Removal: Noise removal involves removing digits of characters or text parts that may interfere with the text analysis. Non-Arabic terms, stop words, numbers, punctuations, emojis, hashtags, and URLs were removed using (NLTK) and regular expression (RE) libraries in Python. Table 6 shows examples of different noise types found within the text and how they are removed.

Table 6: Examples of noise removal

Noise Type	Arabic Tweet Text and Meaning	Classification
Number	yāmryḍ kānt 3 dfʿ māāstlm mnhā tnāzl ʿn dfʿtyn qbl ʿšān āltsǧyl fhmt yā ġby ؟؟	yāmryḍ kānt dfʿ māāstlm mnhā tnāzl ʿn dfʿtyn qbl ʿšān āltsǧyl fhmt yā ġby ؟؟
	You sick man! it was 3 payments I did not receive any payment. Did you get it, idiot	You sick man! it was payments I did not receive any payment. Did you get it, idiot
Usernames	@Sehamwaheed1 wǧhk mrāāāāā mqrf	wǧhk mrāāāāā mqrf
	Your face is disgusting. wrby lāḍrbk yā dbbbbb sāāāmɛʿ	Your face is disgusting. wrby lāḍrbk yā db sāmʿ
Repeated letters	I swear I will hit you fat beaaaar do you heaaar me yrǧāl ṭs lā ʎʿnk ʿbd !!	I swear I will hit you fat bear do you hear me yrǧāl ṭs lā ʎʿnk ʿbd
Punctuations	shut up man damn on you niger!!	shut up man damn on you niger

In the Arabic language, some words contain at least two duplicates letters but they are not redundant, but rather they are part of the word such as the letter “b” in "bbġāʾ," which means "parrot" and the letter "m" in "mmtāz," which means "excellent." These letters should not be removed from the words. Thus, a list of words with redundant letters that are

part of the word itself is created and it contains 35 words. This list is used to check if the redundant letter is a part of the word or not. If the word is included in the list, the redundant letter will not be removed. Table 7 shows some examples of the rendered-letter list.

- Normalization: Text normalization is the process of converting text into one single form by replacing similar letters that are used interchangeably in the Arabic language. Table 8 shows examples of these interchangeable letters. For example, the words (aaḍrbk) and (-aḍrbk), which means "hit you," will be considered as different words by the classifier while they have the same meaning. Additionally, normalizing any diacritics (ḥrkāt) for all the text and removing elongated letters that would appear in the text were also considered. For normalization, the Araby library in python was used. Table 9 shows examples of normalization applied to sentences. By normalizing words, only one form of a word with a specific meaning is used.

- Stop words Removal: Stop-word-removal aims to remove insignificant words. Stop-words are words that are commonly used in a language, and carry no useful information. These words include prepositions, conjunctions, pronouns, and others. While Stop word removal does not affect the meaning of a sentence, it can affect classification performance positively and improve its accuracy. To show the effect of stop word removal, a comparison is done in experiments to compare the performance of the considered model with/without stop words removal. Table 10 shows some examples of the stop word list.

Table 11 shows an example of applying all the pre-processing steps.

3.4. Data classification

The classification is done in two levels as shown in Fig. 3 where the first level is classifying the tweet into violent text or non-violent, then the second level classifies the violent text into cyberbullying or threatening.

The data was classified by using supervised machine learning algorithms namely SVM and NB. Also, two features selection methods were applied. The first one is a pre-trained distributed word representation model named Aravec (Soliman et al., 2017).

Table 7: Examples of the rendered-letter list

Rendered-letter list	Meaning
mmtāz	excellent
bbġā’	parrot
mml	boring
mmtlkāt	possession
tštt	dispersion
mmyz	special
msw--wlytnā	our responsibility

It has been trained on 1,476,715 vocabularies gathered from Twitter. The other method is the term frequency (TF) method, which measures how frequently a term occurs in a document (Utomo and Sibaroni, 2019).

Table 8: Examples of interchangeable letters

Letters	Original Words	Normalized Words	Meaning
-a, a--, aa	*-aḍrbk* *--aḍrbk*	*āḍrbk*	hit you
Y, y--	*ʿly--* *ʿlY*	*ʿly*	on
h’	*dbh’*	*dbh*	fat
w--	*sw--āl*	*swāl*	question

Table 9: Examples of sentence normalization

Normalization Type	Original Sentence	Normalized Sentence
Interchangeable letter such as (h’) and (-a) which are replaced with (h) and (ā).	*lyš ỉnty* *dbh’* Why are you fat *lyšً dًbh* *lyš nḥ◌yf*	*lyš ānty dbh* Why are you fat *lyš dbh* *lyš nḥyf*
Arabic diacritic	Why you are fat.	Why you are fat.
	why you are slim.	why you are slim.
Elongated Letters	*swf āqt.l.k* I will kill you	*swf āqtlk* I will kill you

Table 10: Examples of the stop-word list

Stop-word list	Meaning
ʿlY	on
ālY	to
fy	in
ʿnd	at
mn	from

Table 11: Example of applying all the pre-processing steps

Arabic Sentence				
ūrbīw l’aḍrbk īā dbbbbb sāāām“‘!				
Sentence Meaning				
I swear I will hit you fat beaaaaaar do you heaaaaaaar me!				
After tokenization				
Token 1	Token 2	Token 3	Token 4	Token 5
ūrbīw	*l-aḍrbk*	*yā*	*dbbbbb*	*sāāāmɛʿ*
After removing the noises				
ūrbī	*l-aḍrbk*	*yā*	*db*	*sāmʿ*
After normalizing the tweet				
ūrbī	*lāḍrbk*	*yā*	*db*	*sāmʿ*
After removing the Stop-words				
ūrbī	*lāḍrbk*		*db*	*sāmʿ*

3.5. Model evaluation

For model evaluation, the cross-validation strategy which is a common classifier evaluation strategy divides the dataset randomly into k subsets or "folds" (F1, F2, ..., Fn) of the same size. In the first iteration, the test will be in F1 while the other subset from F2 to Fn are the training data. In the second iteration, F2 will be the test data, and F1, F3, ..., Fn are the train data, and so on (Han et al., 2011). In this

paper, the cross-validation strategy has been applied to the testing set with 10 folds.

In this paper, the cross-validation strategy has been applied to the testing set with 10 folds. Different evaluation measures have been used to evaluate the classification models. These measures are calculated for each test experiment and then averaged over all tests. Assuming the n is the total number of instances in the test dataset, i is an instance tweet in the dataset, and Xi, and Yi are the model predicted and the actual labels, respectively, accuracy, precision, recall, and F-measure can be defined as follows:

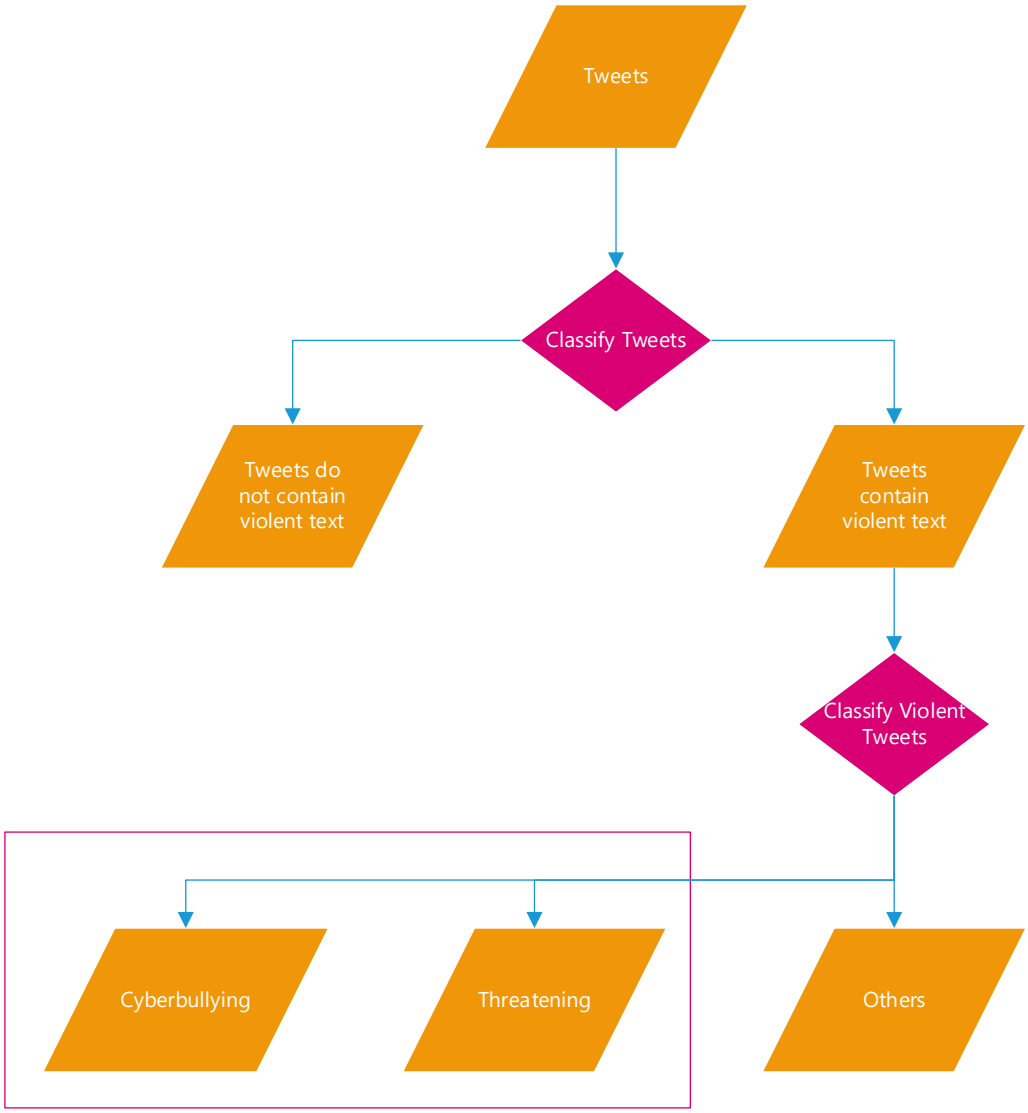

Fig. 3: The suggested system

- Accuracy: is the percentage of all tweets that are correctly classified by the classifier (El-Makky et al., 2014). Eq. 1 is used to calculate the accuracy.

$$A = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i \cap Y_i}{X_i \cup Y_i} \quad (1)$$

- Precision: represents the probability that the tweets which have been classified by the classifier as class X (e.g., cyberbullying) are actually belonging to class X (El-Makky et al., 2014). The following Eq. 2 is used to calculate the precision.

$$P = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i \cap Y_i}{X_i} \quad (2)$$

- Recall: calculates the probability that the tweets of class X (e.g., cyberbullying) are classified as class X by the classifier (El-Makky et al., 2014). Eq. 3 is used to calculate recall as follows:

$$R = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i \cap Y_i}{Y_i} \quad (3)$$

- F-Measure: is an evaluation measure that combines both precision and recall (El-Makky et al., 2014). Eq. 4 is used to calculate F1 as follows:

$$F = \frac{1}{n}\sum_{i=1}^{n}\frac{2(X_i \cap Y_i)}{X_i + Y_i} \quad (4)$$

4. Experiments and results

The following sub-sections describe the experimental setting used in this study along with the results of the conducted experiments.

4.1. The first level of classification

In the first level of classification, the tweets are classified into either violent or non-violent. Four experiments are done at this level. In all of them the tokenization, noise removal, and normalization are applied to the dataset. Also, in all the experiments, a cross-validation strategy is applied. The experiments differ in the feature extraction method, and whether or not stop words are removed. Table 12 shows the details of the experiments. Each experiment is applied using the two algorithms SVM and NB. Table 13 shows the results obtained from the first level of classification.

Table 12: Experiments descriptions

#	Feature extraction method	Removing stop-Word
1	AraVec Model (Soliman et al., 2017)	No
2		Yes
3	Term Frequency TF (Utomo and Sibaroni, 2019)	No
4		Yes

As shown in Table 13, the best percentage achieved by the NB algorithm was obtained in experiment four with the TF method and with stop-word removal. While the best percentage achieved by the SVM algorithm was obtained in experiment three with the TF method and without stop word removal. The highest percentage among all the experiments was in the third experiment using SVM with an accuracy of 73.35% and F1 of 76.06%. Therefore, SVM with the TF method and without stop word removal is selected as the best model at this level. The violent tweets that are classified using this model are used as input for the second level of classification.

Table 13: The results of the first level of classification experiments

Experiment	Evaluation measure	SVM Algorithm	NB Algorithm
1	Accuracy	76.03%	59.65%
	Precision	73.31%	57.86%
	Recall	89.13%	97.45%
	F1	80.37%	72.60%
2	Accuracy	72.34%	59.27%
	Precision	68.91%	57.65%
	Recall	90.75%	96.99%
	F1	78.26%	72.30%
3	Accuracy	73.35%	70.83%
	Precision	75.86%	70.54%
	Recall	76.62%	80.56%
	F1	76.06%	75.13%
4	Accuracy	72.57%	70.43%
	Precision	72.24%	69.55%
	Recall	81.92%	81.94%
	F1	76.61%	75.15%

4.2. The second level of classification

Since the highest results in the first level of classification were obtained without stop word removal, stop words are not removed in the second level. Therefore, two experiments are done using the two different feature extraction methods, which are AraVec pre-trained model and TF. Each experiment is applied using the two algorithms SVM and NB.

The experiment results are shown in Table 14. The highest result was in the second experiment setting which uses the SVM algorithm with an achieved an accuracy of 87.79% and F1 of 89.18%.

Table 14: The results of the second level of classification experiments

Experiment	Evaluation measure	SVM Algorithm	NB Algorithm
1	Accuracy	74.60%	46.23%
	Precision	71.52%	33.33%
	Recall	88.71%	1.73%
	F1	79.07%	3.26%
2	Accuracy	87.79%	71.61%
	Precision	86.51%	81.68%
	Recall	92.21%	61.44%
	F1	89.18%	69.92%

5. Conclusion and future work

This paper presented a model for automatic detection and classification of violent text. This paper aims to create a model that detects the phenomena of cyberbullying and threats in social media using a two-level classifier model. The first level classifies text into violent and non-violent and the second level classifies violent text into cyberbullying and threatening. The dataset was consisting of 2000 tweets collected using Twitter API that was manually labeled. Finally, the tweets are pre-processed to fit into the classifier by removing all the noises. Supervised machine learning was used, the two algorithms SVM and NB were trained in different settings. For the first level of the classification, four experiments were done and the SVM achieves higher percentages using the pre-trained model and stop-word removed. The results were 73.35%, 75.86%, 76.62%, and 76.06% for accuracy, precision, recall, and F1, respectively. In the second level of classification, two experiments were done, SVM achieves higher than NB, using TF with an accuracy of 87.79%, a precision of 86.51%, recall of 92.21%, and F1 of 89.18. Future work will focus on including other types of violent text and adding more features as knowing if the text is considered as a violent text based on the use of emojis, tashkil, and other Arabic dialects.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Alakrot A, Murray L, and Nikolov NS (2018). Towards accurate detection of offensive language in online communication in Arabic. Procedia Computer Science, 142: 315-320. https://doi.org/10.1016/j.procs.2018.10.491

AlHarbi BY, AlHarbi MS, AlZahrani NJ, Alsheail MM, Alshobaili JF, and Ibrahim DM (2019). Automatic cyber bullying detection in Arabic social media. International Journal Engineering Research and Technology, 12(12): 2330-2335.

Al-Kabi MN, Al-Qwaqenah AA, Gigieh AH, Alsmearat K, Al-Ayyoub M, and Alsmadi IM (2016). Building a standard dataset for Arabia sentiment analysis: Identifying potential annotation pitfalls. In the IEEE/ACS 13th International Conference of Computer Systems and Applications, IEEE, Agadir, Morocco: 1-6. https://doi.org/10.1109/AICCSA.2016.7945822

Al-Khalifa H, Magdy W, Darwish K, Elsayed T, and Mubarak H (2020). Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection. In 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France.

Alruily M (2020). Issues of dialectal Saudi twitter corpus. International Arab Journal of Information Technology, 17(3): 367-374. https://doi.org/10.34028/iajit/17/3/10

Altaher A (2017). Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting. International Journal of Advanced and Applied Sciences, 4(8): 43-49. https://doi.org/10.21833/ijaas.2017.08.007

Biltawi M, Al-Naymat G, and Tedmori S (2017). Arabic sentiment classification: A hybrid approach. In the International Conference on New Trends in Computing Sciences, IEEE, Amman, Jordan: 104-108. https://doi.org/10.1109/ICTCS.2017.24

Díaz-Torres MJ, Morán-Méndez PA, Villasenor-Pineda L, Montes M, Aguilera J, and Meneses-Lerín L (2020). Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association, Marseille, France: 132-136.

Duwairi RM and Qarqaz I (2014). Arabic sentiment analysis using supervised classification. In the International Conference on Future Internet of Things and Cloud, IEEE, Barcelona, Spain: 579-583. https://doi.org/10.1109/FiCloud.2014.100

El-Makky N, Nagi K, El-Ebshihy A, Apady E, Hafez O, Mostafa S, and Ibrahim S (2014). Sentiment analysis of colloquial Arabic tweets. In the ASE BigData/SocialInformatics/PASSAT/ BioMedCom 2014 Conference, Harvard University, Cambridge, USA: 1-9.

El-Naggar N, El-Sonbaty Y, and Abou El-Nasr M (2017). Sentiment analysis of modern standard Arabic and Egyptian dialectal Arabic tweets. In the Computing Conference, IEEE, London, UK: 880-887. https://doi.org/10.1109/SAI.2017.8252198

Gambäck B and Sikdar UK (2017). Using convolutional neural networks to classify hate-speech. In the 1st Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, Canada: 85-90. https://doi.org/10.18653/v1/W17-3013

Haidar B, Chamoun M, and Serhrouchni A (2017). Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In the 1st Cyber Security in Networking Conference, IEEE, Rio de Janeiro, Brazil: 1-8. https://doi.org/10.1109/CSNET.2017.8242005

Haidar B, Chamoun M, and Yamout F (2016). Cyberbullying detection: A survey on multilingual techniques. In the European Modelling Symposium, IEEE, Pisa, Italy: 165-171. https://doi.org/10.1109/EMS.2016.037

Han J, Kamber M, and Pei J (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4): 83-124. https://doi.org/10.1016/B978-0-12-381479-1.00003-4

Husain F (2020). Arabic offensive language detection using machine learning and ensemble machine learning approaches. Available online at: https://arxiv.org/abs/2005.08946

Ismail R, Omer M, Tabir M, Mahadi N, and Amin I (2018). Sentiment analysis for Arabic dialect using supervised learning. In the International Conference on Computer, Control, Electrical, and Electronics Engineering, IEEE, Khartoum, Sudan: 1-6. https://doi.org/10.1109/ICCCEEE.2018.8515862

Kumar R, Ojha AK, Malmasi S, and Zampieri M (2018). Benchmarking aggression identification in social media. In the First Workshop on Trolling, Aggression and Cyberbullying, Santa Fe, USA: 1-11.

Mahmud A, Ahmed KZ, and Khan M (2008). Detecting flames and insults in text. In the 6th International Conference on Natural Language Processing, CDAC Pune, Pune, India.

Modha S, Majumder P, and Mandl T (2018). Filtering aggression from the multilingual social media feed. In the 1st Workshop on Trolling, Aggression and Cyberbullying, Santa Fe, USA: 199-207.

Mohaouchane H, Mourhir A, and Nikolov NS (2019). Detecting offensive language on Arabic social media using deep learning. In the 6th International Conference on Social Networks Analysis, Management and Security, IEEE, Granada, Spain: 466-471. https://doi.org/10.1109/SNAMS.2019.8931839

Mouheb D, Abushamleh MH, Abushamleh MH, Al Aghbari Z, and Kamel I (2019). Real-time detection of cyberbullying in Arabic Twitter streams. In the 10th IFIP International Conference on New Technologies, Mobility and Security, IEEE, Canary Islands, Spain: 1-5. https://doi.org/10.1109/NTMS.2019.8763808

Mouheb D, Ismail R, Al Qaraghuli S, Al Aghbari Z, and Kamel I (2018). Detection of offensive messages in Arabic social media communications. In the International Conference on Innovations in Information Technology, IEEE, Al Ain, UAE: 24-29. https://doi.org/10.1109/INNOVATIONS.2018.8606030

Mubarak H, Darwish K, and Magdy W (2017). Abusive language detection on Arabic social media. In the 1st Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, Canada: 52-56. https://doi.org/10.18653/v1/W17-3008

Mubarak H, Rashed A, Darwish K, Samih Y, and Abdelali A (2020). Arabic offensive language on Twitter: Analysis and experiments. Available online at: https://arxiv.org/abs/2004.02192

Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, and Wojatzki M (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. Available online at: https://arxiv.org/abs/1701.08118

Schneider JM, Roller R, Bourgonje P, Hegele S, and Rehm G (2018). Towards the automatic classification of offensive language and related phenomena in German tweets. In the 14th Conference on Natural Language Processing KONVENS, Viena, Austria: 95-103.

Soliman AB, Eissa K, and El-Beltagy SR (2017). Aravec: A set of Arabic word embedding models for use in Arabic NLP. Procedia Computer Science, 117: 256-265. https://doi.org/10.1016/j.procs.2017.10.117

Spertus E (1997). Smokey: Automatic recognition of hostile messages. In the 9th Conference on Innovative Application of Artificial Intelligence, Providence, USA: 1058-1065.

Utomo MRA and Sibaroni Y (2019). Text classification of British English and American English using support vector machine. In the 7th International Conference on Information and Communication Technology, IEEE, Kuala Lumpur, Malaysia: 1-6. https://doi.org/10.1109/ICoICT.2019.8835256

Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, and Hoste V (2015). Detection and fine-grained classification of cyberbullying events. In the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria: 672-680.

Vieira SM, Kaymak U, and Sousa JM (2010). Cohen's kappa coefficient as a performance measure for feature selection. In the International Conference on Fuzzy Systems, IEEE,

Barcelona, Spain: 1-8. https://doi.org/10.1109/FUZZY.2010.5584447

Wiegand M, Siegel M, and Ruppenhofer J (2018). Overview of the germeval 2018 shared task on the identification of offensive language. In the 14th Conference on Natural Language Processing, Austrian Academy of Sciences, Vienna, Austria: 1-10.