Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach

Monther Khalafat
The University of Jordan, Amman, Jordan

Ja'far S. Alqatawna
The University of Jordan, Amman, Jordan
Higher Colleges of Technology, Dubai, United Arab Emirates

Rizik Al-Sayyed (✉), Mohammad Eshtay
The University of Jordan, Amman, Jordan
r.alsayyed@ju.edu.jo

Thaeer Kobbaey
Higher Colleges of Technology, Dubai, United Arab Emirates

Abstract—Today, the influence of the social media on different aspects of our lives is increasing, many scholars from various disciplines and majors looking at the social media networks as the ongoing revolution. In Social media networks, many bonds and connections can be established whether being direct or indirect ties. In fact, Social networks are used not only by people but also by companies. People usually create their own profiles and join communities to discuss different common issues that they have interest in. On the other hand, companies also can create their virtual presence on the social media networks to benefit from this media to understand the customers and gather richer information about them. With all of the benefits and advantages of social media networks, they should not always be seen as a safe place for communicating, sharing information and ideas, and establishing virtual communities. These information and ideas could carry with them hatred speeches that must be detected to avoid raising violence. Therefore, web content mining can be used to handle this issue. Web content mining is gaining more concern because of its importance for many businesses and institutions. Sentiment Analysis (SA) is an important sub-area of web content mining. The purpose of SA is to determine the overall sentiment attitude of writer towards a specific entity and classify these opinions automatically. There are two main approaches to build systems of sentiment analysis: the machine learning approach and the lexicon-based approach. This research presents the design and implementation for violence detection over social media using machine learning approach. Our system works on Jordanian Arabic dialect instead of Modern Standard Arabic (MSA). The data was collected from two popular social media websites (Facebook, Twitter) and has used native speakers to annotate the data. Moreover, different preprocessing techniques have been used to show their effect on our model accuracy. The Arabic lexicon was used for generating feature vectors and separate them to features set. Here, we have three well known machine learning algorithms: Support Vector Machine (SVM), Naive Bayes (NB) and k-Nearest Neighbors

(KNN). Building on this view, Information Science Research Institute's (ISRI) stemming and stop word file as a result of preprocessing were used to extract the features. Indeed, several features have been extracted; however, using the SVM classifier reveals that unigram and features extracted from lexicon are characterized by the highest accuracy to detect violence.

Keywords—Violence Detection, Social Networks, Sentiment Analysis, Arabic Lexicon, Feature Extraction

1 Introduction

Nowadays, with the existence of the internet in life, people began creating communities online. Social media networks have become an essential part of the Internet for billions of people all over the world. These networks are equally important for both people and organizations over the globe. The information and photo sharing, spreading ideas, exchange experience and recommendations are attractive properties of social networks. They enable clients to express their opinions, show loyalty and start conversations with their favorite companies. On the other hand, companies use social media networks to collect information about markets, clients, and competitors in an easy way. Furthermore, companies can communicate with clients to enhance their image and meet their expectations. Based on this view, it is critical to build a trust in using the social media for both people and companies. In fact, building trust in social websites plays a critical role in enhancing the quality of social media networks and implementing security for them. One of the key aspects when dealing with social networks is to try to minimize the effect of hate speech which may lead to violence, terrorism, and security problem to the political systems [1]. One of the most popular methods that help in detecting certain types of semantics in a data is Sentiment Analysis (SA) [2].

SA is a way for text mining, computational treatment of opinions, sentiments and subjectivity of the text [3]. SA is a field of study that measures people's opinions, sentiments through natural language processing (NLP), computational linguistics and text analysis that are used to extract and analyze subjective information from the web, mostly social media and similar sources [4].

In this study, we are concerned with the activity of Jordanian people in the major social networks. According to Arabic social media report, there were 156 million Facebook users in the Arab world of whom 5 million in Jordan and 11.1 million Twitter users of whom 200 thousand in Jordan in March 2017 [5]. In Facebook and Twitter, people share their likes, dislikes, beliefs, political and sports opinions. Among these opinions there are a significant percentage of violence and abuse statements.

Sites such as Facebook and Twitter have become a priority to actively combat hate speech which leads to violence [6]. Some of these sites such as Facebook added options to report violent in the feedback options for any post. The importance of detecting hate speech is clear from the strong relation between hate speech and actual vio-

lence. Early detecting hate speech could enable outreach programs that attempt to prevent an escalation from speech to action.

In this research, we built a model for violence detecting by using people (tweets, posts) written in Jordan Arabic dialect through popular social media networks sites (Twitter, Facebook).

Despite the increasingly massive number of Arabic users on the Internet [5], Arabic language is considered amidst top six major languages of the world. The number of native speakers exceeds 200 million and it is the formal language used in over twenty countries [7], there is a weakness for building a strong corporation to exploit it in different applications. There are three different forms of Arabic language [2]: Modern Standard Arabic (MSA), Dialectal Arabic (DA) and Classical Arabic.

This study deals with DA, but still, there are many challenges when using Arabic language as listed below:

1. Colloquial Arabic parsers: Many people used their Dialectal language instead of MSA on social media, parsing MSA is an already complex task towards which many efforts have been directed. However, colloquial Arabic different from MSA phonological, morphological such as ("walad", which means "a boy") ("waldan", which means "two boys"), and ("awlad", which means "more than two boys" [7], and lexical and does not have a standard orthography which complicates the task of building morphological analyzers and part of speech taggers [8].
2. Pronunciation: Some pronunciations do not exist in English such as "Gh" as in "Gharb" [7].
3. Diacritics: such as, "teacher" "مُدَرِّسة" and "school" "مَدْرسَة" [9].
4. Poor of sentiment lexicons: Sentiment lexicons contain opinions with their polarity, they are an important part of any sentiment analysis. There are currently a few number of publicly available colloquial Arabic sentiment lexicons, so building a colloquial Arabic polarity lexicon is still an open research area.
5. Named entity recognition: Named entity recognition becomes an important part of sentiment analysis when identifying the polarity of the opinion. Person name recognition becomes a requirement even for the task of determining semantic orientation, such as "نبيل, سعيد ".
6. Phrases and idioms: phrases and idioms are very commonly used by Arabic speakers in social media to express their opinions and feelings in a sarcastic way, old wisdom and idioms, consequently, we shall exclude them because it is hard to deal with them.
7. Negation: Negation can be an important concern in opinion and sentiment analysis. For example, "أنا أحب هذا الكتاب" "أنا لا أحب هذا الكتاب". There is a list of negation terms in Arabic language which can change the sentiment polarity of terms from negative to positive and vice versa.
8. Emoticons: Arabic smiles and sad emoticons are often mistakenly interchanged; so many tweets have words and emoticons that are contradictory in sentiment mainly due to mixing the text orientation while typing emoticons.

As mentioned above, DA is the target of our study so we have collected the dataset manually and it has been annotated into (Violence, Normal) by Arabic Jordan native speakers.

In this study, we have built our own dataset that consists of set of words that were extracted from Facebook posts and twitter tweets in the Jordanian dialect. The collected dataset is then handled by applying set of pre-processing techniques to enhance the generalization performance of violence detection. This manipulated dataset and the lexicon are used as a source for the extracted features. In order to detect violence using the extracted features, three well-known classification algorithms are used (Naïve Bayes, Support Vector Machines and K-Nearest Neighbour)

Many experiments have been conducted to study the effect of applying various pre-processing techniques and different classifiers. The results show that SVM classifier is doing better than KNN and NB when used with the collected dataset

This paper is organized as follows. There are five more sections. In section 2, Related work presents an overview of related work. In section 3, Proposed method will be discussed Under this section data collection and pre-processing, building lexicon, features extraction, machine learning approach will be discussed. Section 4 discusses Experimental Results. Finally, in section 5, conclusion and future work.

2 Related Work

Millions of users share opinions on different aspects of life every day. Therefore, social media web-sites like (Facebook, Twitter) are rich sources of data for opinion mining and sentiment analysis [24]. One of the application of this mining and analysis process is the detection of violent and hate speech. Different approaches were proposed in the literature to tackle the problem of sentiment analysis.

Hammer (2014) [10] presented a method of using machine learning to detect threats of violence from a data set of YouTube comments written in English language. The method described in the paper uses logistic LASSO regression analysis on bigrams of important words to classify sentences as violence or not. The dataset contains 24840 sentences from YouTube and was manually annotated as violent threat or not. The features are bigrams of two of these important words observed in the same sentence. The paper did not describe properly how these important words were selected, report only that words were chosen that were correlated with the response (violent/non-violent). However, it appears likely that the words were arrived at using LASSO regression. The obtained result shows Accuracy of 0.9466. The shortcoming of this study was the using of the logistic LASSO regression analysis, which has the limitation of performing an implicit feature selection while estimating the model.

In another work, Djuric et al. [6] proposed a method to detect hate speech comments in English language by using two steps. First, they used paragraph2vec to convert a generic block of text into a vector using continuous bag of words (CBOW) which used for predicting the word given its context. Second, they used logistic regression classifier. The proposed model compared to TF and TF-IDF using area under

curve (AUC) metric. The results show that Paragraph2vec outperforms TF and TF-idf.

In their paper, Yadav and Manwatkar (2015) [11] developed a social media networks prototype that aims to automatic filtering of offensive content in social media networks before sharing it. They applied AHO-Corasick string pattern matching algorithm. The idea of the proposed algorithm is based on matching the pattern (i.e. offensive keywords) from the input text with database that have offensive keywords collected from different datasets. After the word is detected they simply replaced it with some special character to prevent it from share. They used breadth first search to find the offensive keywords. The prototype shows good results even when using slang language.

Gitari et al. (2015) [12] created a model classifier that uses sentiment analysis techniques and in particular a lexicon-based approach to automatically detect hate speech in online forums, blogs and comments in news reviews using semantic and subjectivity features. They collected blogs from Raymond Franklin sites that are considered to be generally offensive, they refer to this as first corpus. Second corpus consists of largely paragraph related to the Israel-Palestinian conflict. They concentrated classifying hate speech detection into three key target groups of race, nationality and religion. The achieved results show that precision, recall, and F-score are the best when using semantic orientation, hate verbs and theme-based as feature sets in first corpus and using subjective sentences only, while the results much less when used same feature sets but without using subjective sentences. From my opinion they could increase precision and recall scores if they applied machine learning.

Waseem and Hovey (2016) [13] presented a method for hate speech detection on Twitter that is written in English. The method is divided into three parts. First, they collected 16K tweets that contain racist and sexist hate speech, then they proposed a list to identify hate speech which helps the annotators to reliably identify hate speech. Second, they examined the impact of different extra-linguistic features in coupling with character n-grams for hate speech detection instead of using word n-gram due to character n-gram being far less sparse than the word n-gram. In third part, they used a grid search in order to select the most suitable features, after that they evaluated the selection features by using logistic regression classifier and 10-fold cross validation. The obtained results show that using character n-grams of length up to 4 with gender as an additional feature have got 73.93 F1-score, which it is the best compared with location and word n-grams. The limitation in this article centred in extracted location feature which it need to consider more than just the tags Twitter provides, which affected on the final result.

Alhelbawy et al. (2016) [14] presented a new corpus of violent tweets event; the dataset was manually labeled for seven classes of violence (Human rights abuse, Political opinion, Accident, Crime, Conflict, Crisis, Other) and annotated using popular crowdsourcing platform crowdflower. The work was targeting the Arabic tweets that are related to violence. Two filters are then applied to filter (Redundant tweets, emotional tweets, delete short tweets and sexual adverts), after that they used the confidence score to evaluate different sub-sets for each tweets. The obtained results show

that crowd classification is overall reliable and can be used for further research on violence on social media.

Mubarak et al. (2017) [15] presented an automated method to create and expand a list of obscene words that help to detect abusive language on Arabic social media. They classify Twitter users based on the using of abusive words or not. By using Twitter API with language filter set to Arabic they collected tweets patterns that are usually used in offensive communications. After that, they manually considered if the words that appeared in collected tweets are obscene or not.

Magu et al. (2017) [1] introduced a mechanism to identify hate coded content on social media written in English language. For example, user have used words skype, google to represent Jew, black respectively. This hate coded leads to violence over social media. They extracted 1999 tweets separated to 1048 labeled hateful and the rest were labeled non-hateful. They used dataset to calculate the Pearson correlation coefficient appearance of every term exist in the dataset and the class label, this method helps them to extract the most correlated terms for identifying hate tweets. After that, they ran the Apriori algorithm to extract frequent item set. Lastly, to train classifier they applied a bag of words model to represent the most popular word in the corpora of training dataset with a Boolean feature vector, then they collected 23,401 tweets to build model using support vector machine with linear kernel to be able separate the hateful tweets and using 10-fold cross validation. They achieved an accuracy of 0.794 with precision of 0.794 and recall of 0.795.

Abdelfatah et al. (2017) [16] proposed a new framework aiming at separating violent and non-violent Arabic tweets over Twitter by using sparse Gaussian process latent variable model (SGPLVM) followed by k-means. Experiment started by collecting 16234 Arabic tweets then pre-processing tweets by removing stop word and web links, after that they annotated manually by at least five different annotators and only tweets have a confidence score more than 0.7 are applied, then two set of experiments have been carried out. The first one is reduced dimensions with PCA and then apply K-means. Second experiment is using SGPLVM to reduce dimension then apply K-means, after that they compare between two results. The obtained results show that using SGPLVM with K-means it better than PCA with K-means and using unsupervised techniques to detect violent tweets in low dimensional representation be better than applying clustering on the original data.

3 Proposed Method

This section describes in details the methodology adapted in this research. The main processes in the research are: Data collection, Data preprocessing, Arabic lexicon, Feature extraction, Data classification and Model evaluation. Figure 1 illustrates the various processing steps of the proposed method.

3.1 Data collection

Data collection is the process of gathering large amount of data, in an established systematic fashion that enables one to prove research questions, test hypotheses, and evaluate results. In order to test our model, we need to use some datasets that contain data from Arabic audience. There are few datasets that contain this information but they are not useful in our research since they are collected in Arabic delicts that are different from our target delict which is the Jordanian Arabic delicts. In order to overcome this situation, we have built dataset from scratch that contain Jordanian dialect opinions about subjects related to politics and sports. The created dataset contains set of comments collected in various times from two popular social networks – Facebook and twitter [25]. These comments are political and sport comments that might lead to violence. Table 1 shows example of the collected data divided into violence and normal data.

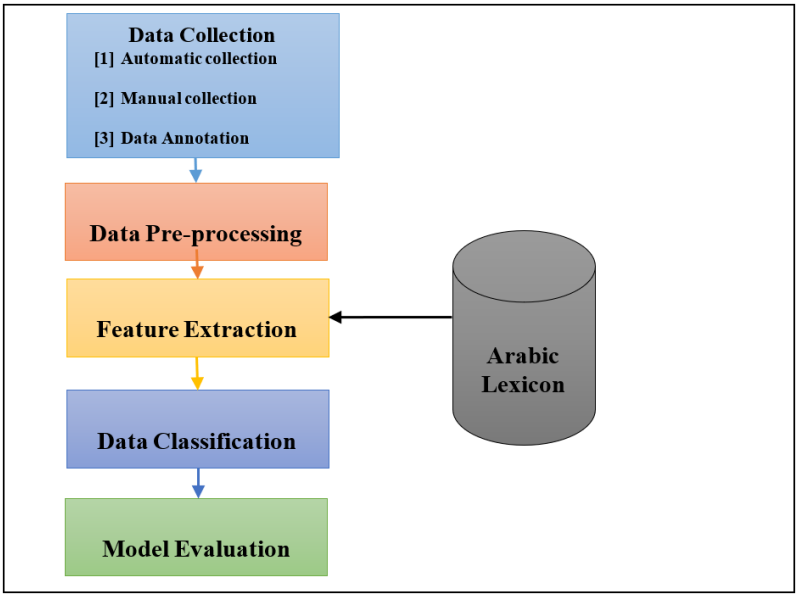

Fig. 1. Research Processes.

Table 1. Example about collected data

Violence Data	Normal Data
الحرامي حرامي من دار ابوه والله لو معه مليار ليرة بده يسرق	اللهم ارزق الاردن ناس قلبها نظيف وبتخاف الله
صارت مسخرة كل نادي يجيب حكام هو دافعلهم الاخطاء والظلم التحكيمي ضيعت متعة كرة القدم اتمنى يجيبو حكام من الصين لا ليهم في العير ولا النفير يحكمو بحق الله و بس	الفيصلي يعود أمام الجزيرة ويحافظ على صدارة
أعمال شغب واسعة في منطقة الخيرية بمحافظة اربد	ملكنا الله يقويك ويعينك على كل هالمكائد التي تكاد حول البلد من الخارج والداخل
لعنة الله على هيك حكومه تدمر شعبها	اخي الحبيب الفيصلي والوحدات إخوة وهم يشكلون المنتخب الوطني لكرة القدم العزيز

3.2 Data collection methods

In this study, two methods were used to collect data from Facebook and Twitter: automatic and manual collection. In the first method, we used scripts to collect data automatically from the tweets of twitter. The number of tweets collected is almost 10,330 tweets. The number of useful tweets after excluding redundant and advertisement tweets is 385, 335 among them are violence and the rest is regular. The automatic collection of data was not enough, so we used manual collection. The main reasons for adopting this method despite the time and effort needed are:

- Most of tweets returned from automatic collection are violence and it is hard to find keyword not including violence content.
- There is no automatic data collection for Facebook.
- Huge number of returned tweets are advertisement, links and comments on picture.

Three persons that are aware of Jordanian dialect, politics and sport performed the data collection. They revised around 60000 tweets and comments. The final number of collected tweets and comments is 2057. Table 2. shows source, the domain and the number of followers of the groups that were used to collect data from.

Table 2. Groups that collected data from them and number of followers for each of them.

Social media name	Name of group	Domain	Number of followers
Facebook	Khaberni	Politics	>2M
Facebook	Almnthb alardny	Sport	>150K
Twitter	Alghad	politics	>700K
Twitter	Kooora	sport	>1M

The next step after data collection is data annotation. Each tweet or comment should be classified as violence or normal. Table 3 summarizes the results of data annotations.

Table 3. Number of violence\normal data after annotation.

Violence	Normal	Total number
1155	902	2057

3.3 Data preprocessing

In the pre-processing stage, various Natural Language Processing (NLP) techniques are applied [9]. Several preprocessing strategies can be applied on sentiment analysis that affect the accuracy when applied to Arabic text. The pre-processing is performed on different stages: Tokenization, normalization, stop-word removal, and stemming.

In this research, we applied different pre-processing techniques. At the beginning, tokenization is used to break up the text into tokens. Next, normalization is applied to convert all various forms of a word to a common form. It is the process of transform-

ing text into a single canonical form since input is guaranteed to be consistent before operations are performed on it [17]. In this study, normalizer performs this specific task according to the rules listed below:

- Replacing: in this stage we want to replace any words contain characters " أ " replace by " ا " , "ة" replace by "ه" .
- Removing the "tatweel" character " _ ".
- Removing the diacritics: "يَشْرَبُ المَاءَ " will be "يشرب الماء".
- Removing punctuation and special characters.
- Remove English letters and numbers.
- Remove hyperlink: some of collected data contain link such as for continue reading the news, in this situation we deleted the link and keep the news text.
- Remove duplicate letters such as "مبرووووووك" which should be "مبروك".

The normalized text is then handled in order to remove stop words. These words are removed because they don't add any new information to the text. A list of stop words such as"كذلك, هو, انت " is prepared and applied to the text. The last process applied is stemming. In stemming, all affixes of the words are removed. For instance, the word "المهاجرون" is stemmed into "مهاجر". Light stemming python library used for this purpose. It provides a configurable stemmer and segmented for Arabic text (Zerrouki,2012) . Table 4, shows some examples of applying Light Stemming on original data.

Table 4. Example of applying Light Stemming.

Original Data	Light Stemming
راح تقلب الأردن شيكاغو من وراء الجوع والقهر	راح تقلب اردن شيكاغو من وراء جوع قهر
نشامى الله يعطيهم الصحه والعافيه قولو امين	نشامى الله يعطي صحه عافيه قولو امين
اكيد امن الاردن او بنفتخر بهاد الكلام	اكيد امن اردن او بنفتخر بهاد كلام
يجب ان نحترم التشجيع وانا لا نتطرق الى امور اخرى	يجب ان نحترم تشجيع وانا لا نتطرق الى امور اخرى

3.4 Normalizing the dataset

In this phase, after extracting the features, the dataset is normalized. Normalization is generally performed during the data preprocessing. Since the scaling of values of dataset attributes range varies widely so it helps to fit into specific range. There are different normalization types such as Z-score and Min-Max normalization [18]. In this research Min-Max normalization type is used to scale data into range [0,1] because the dataset features have different range such as the range of feature the total frequency of violence words is [0,6]. The following equation is the formula of Min-Max normalization to scale the data in range [0,1].

$$v' = \frac{v - min_A}{max_A - min_A} \quad (1)$$

Where v refers to original value and v' refers to new value.

3.5 Arabic lexicon

Many researches work in the problem of sentiment analysis use sentiment lexicons also known as senti-lexicons. In this research we relied on Arabic senti-lexicon (ASL) that proposed by [19] where each word is assigned a polarity (positive, negative). The sentiment lexicon is considered as the most crucial resource for features extraction that help us to make model more accurate. ASL includes 13,760 positive and negative words, 3880 synset terms which have different words and same meaning that collected manually and dialects synst(D-synset) which are words used in the different Arabic dialects and have the same meaning.

Based on ASL, we built our own lexicon to detect violence. Firstly, ASL was modified to become appropriate for violence detection where negative is changed to normal and positive is changed to violence. Secondly, 304 violence words and 225 normal words were added manually to ASL. Finally, we deleted D-synset because we apply Jordanian Dialect and we deleted the score field. In addition, we added set of part of speech (POS) and synset and inflection manually.

Our new lexicon includes a list of 10,443 violence and normal words and 2450 violence and normal sentiment. An example from our new lexicon is shown in table 5.

Table 5. Some example from our new lexicon.

Word	POS	Polarity	Synset	Inflection terms
نشامى	Adj	Normal	-	نشميات
تحقر	Verb	Violence	هان;ذل;حقر;ضعف;صغر;احس;وضع;اذل	حقيرون,حقيرات
وسخ	Noun	Violence	وسخ;تافر;دنس;رجس;قذر	وسخين

3.6 Features extraction

Feature extraction is an important task in the sentiment analysis and more generally in text categorization. Since the text is unstructured, we need to convert original documents into feature vectors which is the main step in any supervised learning approach attached to sentiment analysis to select the right features that determine the overall performance of sentiment classification. Consequently, this work studies the effect of applying various pre-processing techniques on the extracted set of features and its impact on the overall performance.

In this research, we define a group of features. These features can be grouped into four main feature groups:

1. Feature based on sentiment word of (Violence, Normal), presence and frequency. In this group, we defined six features: The presence of violence words, the total frequency of violence words, the presence of normal words, the total frequency of normal words, the ratio of the presence of violence to normal words, and the ratio of the total frequency of violence to normal words.
2. Bag-Of-Words (BOW) [20]: According to the BOW model, the document is represented as a vector of words in Euclidian space where each word is independent from others [21]. We used the feature n-grams and its Term Frequency–Inverse

Document Frequency (TF/IDF). The n-grams of texts are used in text mining and natural language processing tasks. N-grams are simply all combinations of adjacent words or letters of length N that you can find in text. An n-gram of (N= 1) is referred to "unigram"; (N=2) is a "bigram": N= 3 is a "trigram". In this research we used the unigram, bigram and trigram.

3. Features based on POS: The process of POS tagging allows to tag each word of text in terms of which part of speech it belongs to: noun adjective, verb, etc. The goal is to extract patterns in text based on analysis of frequency distributions of these part-of-speech. There is no common opinion about whether POS tagging improves the results of sentiment classification. Barbosa and Feng (2010) [22] reported positive results using POS tagging, while Kouloumpis et al. (2011) [23] reported a decrease in performance, so it depends on domain that you work in. We defined nine features: total number of violence adjective, total number of violence verb, total number of violence noun, total number of normal adjective, total number of normal verb, total number of normal noun, the ratio of the total frequency of violence adjective to normal words, the ratio of the total frequency of violence verb to normal words, and the ratio of the total frequency of violence noun to normal words.
4. Other features: We defined two features: the presence of negation word in sentence and the presence of violence and normal emoticons in sentence.

Table 6 represents the summary of features extracted from (Tweet, Post).

Table 6. Features extracted from each (Tweet, Post)

Feature group name	Feature number and name
Sentiment words of presence and frequency	F1: The presence of violence word F2: The total frequency of violence words F3: The presence of normal words F4: The total frequency of normal words F5: The ratio of the presence of violence to normal words F6: the ratio of the total frequency of violence to normal words
Bag-Of-Words (BOW)	F7: The TF/IDF of unigram, bigram and trigram independently
Features based on POS	F8: the ratio of the total frequency of violence adjective to normal words F9: the ratio of the total frequency of violence verb to normal words F10: the ratio of the total frequency of violence noun to normal words F11: total number of violence adjective F12: total number of violence verb F13: total number of violence noun F14: total number of normal adjective F15: total number of normal verb F16: total number of normal noun
Other features	F17: presence of negation word in sentence F18: presence of violence emoticons in sentence F19: presence of normal emoticons in sentence

Another result from the models to measure is the AUC (area under curve), to calculate this measure values using Python Sklearn matrix, the model should be first

trained and tested using the test data, then the evaluation methods can be applied to calculate the final results (which are presented in the experiment results section). These metrics of AUC and ACC are commonly used in the evaluation of the link prediction problems. After the model building and evaluation phases, the main results from this research have to be evaluated by comparing the results from the different models.

There are two ways to compare these results. The first traditional one is by having that comparison manually using the resulted values from the described evaluation matrices. The second approach is more scientific method to compare the whole models' accuracy using the statistical significance tests and then apply the AUC after having confidence in the differences between models.

4 Experiments

In this research, three classifier methods were used to detect violence in Arabic language using sentiment analysis; SVM, NB and KNN. These methods used due to their effectiveness, simplicity and accurateness.

4.1 Parameters of classification algorithms

Generally, many machine learning algorithms need set of parameters to be assigned. Table 7 lists the algorithms used, the parameters of classification algorithms as well as the selected values.

Table 7. parameters of classification algorithms

Algorithm	Parameter	Value
SVM	kernel gamma degree C	RBF Auto 3 1.0
NB(MultinomialNB)	alpha class_prior fit_prior	1.0 None True
KNN	leaf_size metric k	30 Euclidean Distance 5

4.2 Model evaluation

To evaluate the quality and usefulness of the model, several experiments were conducted on our dataset. All algorithms were evaluated using 10-fold cross validation. The measures used to evaluate the models are based on the confusion matrix depicted on Table 8.

Table 8. Confusion Matrix

	P' (Predicted)	**N' (Predicted)**
P (Actual)	True Positive (TP)	False Negative (FN)
N (Actual)	False Positive (FP)	True Negative (TN)

In most of sentiment analysis problems, three measures are used to evaluate the model: Accuracy, precision and recall, in addition we used f-measure to measure the accuracy of test data as it considers both precision and recall. The following describe these measures:

- Accuracy: Tthe proportion of the total number of predictions that were correct classified.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

- Recall (R): Tthe proportion of the number of correct positive predictions divided by the total number of positives.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

- Precision (P): the proportion of the number of correct positive predictions divided by the total number of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

- F-Measure: is the weighted harmonic mean of precision and recall.

$$\text{F-Measure} = \frac{2*R*P}{(R+P)} \tag{6}$$

In order to evaluate the robustness of the classifier, the standard deviation (Stdv) of the 10-folds is calculated and reported. Classifiers with lower Stdv show more robustness. The formula of Stdv is listed below:

$$\sigma = \sqrt{\frac{\Sigma\,(X-\bar{X})^2}{n-1}} \tag{7}$$

where n is the number of data points,$\bar{x}$ is the mean of X_i and X_i is each of the values of the data.

4.3 Experimental results

This section provides the detailed experiment results. The purpose of this experiment is to evaluate the violence detection over social media by developing language resources for Arabic sentiment analysis. It lists the results of features that extracted from sentence without using Arabic lexicon such as n-gram and features extracted using Arabic lexicon such as number of violence words in sentence, then a compari-

son of the results for various preprocessing techniques that were used with the features extracted.

All experiments have been performed using Python 2.7, Anaconda Spyder 3.2.3. The experiments have been conducted in a computer with Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz 2.60GHz running Windows 10 64-bits. The computer contains 32 GB RAM.

4.4 Results based on n-gram feature with various preprocessing files

This section represents the comparison results between the different types of n-gram features (unigram, bigram and trigram) on various preprocessing techniques by using three classifiers, SVM, NB and KNN. In order to evaluate n-gram feature, four measures used: recall, precision, accuracy and f-measure. Table 9 shows the results of n-gram feature. In first column is the file that result after applying the following preprocessing techniques:

1. Normalization file: This file contains data that has been applied to normalization without stemming and remove stop words.
2. Stop words file: This file contains data that has been applied to normalization and remove stop words.
3. ISRI stemming file: This file contains data that has been applied to normalization and ISRI stemming.
4. Light stemming file: This file contains data that has been applied to normalization and light stemming.
5. ISRI stemming and stop words file: This file contains data that has been applied to normalization, remove stop words and ISRI stemming.
6. Light stemming and stop words file: This file contains data that has been applied to normalization, remove stop words and light stemming.

We can notice that, NB performed better than the others on normalization file with bigram, stop words with unigram, light stemming with bigram. On the other hand, SVM surpass others on ISRI stemming file unigram feature and ISRI stemming with stop words file bigram feature. Overall, the best results recorded using SVM on ISRI stemming with stop words file with bigram feature. The least result was when using KNN classifier with trigram feature on ISRI stemming and stop word file.

Table 9. Results based of n-gram feature

File name	Measures	Unigram			Bigram			Trigram		
		SVM	***NB***	***KNN***	***SVM***	***NB***	***KNN***	***SVM***	***NB***	***KNN***
Normalization	Recall	0.74	0.83	0.59	0.75	0.84	0.64	0.75	0.84	0.61
	precision	0.69	0.70	0.60	0.70	0.70	0.63	0.69	0.70	0.63
	Accuracy	0.67	0.71	0.56	0.68	0.71	0.59	0.68	0.71	0.59
	F-measure	0.72	0.76	0.60	0.72	0.76	0.63	0.72	0.76	0.62
	Stdev	0.06	0.06	0.07	0.06	0.06	0.06	0.06	0.07	0.04
Stop words	Recall	0.74	0.84	0.73	0.78	0.83	0.80	0.77	0.83	0.82
	precision	0.69	0.70	0.62	0.70	0.71	0.64	0.70	0.71	0.64
	Accuracy	0.67	0.71	0.66	0.69	0.72	0.64	0.69	0.72	0.64
	F-measure	0.71	0.76	0.67	0.74	0.76	0.71	0.73	0.76	0.72
	Stdev	0.06	0.08	0.07	0.09	0.07	0.09	0.04	0.05	0.04
ISRI stemming	Recall	0.84	0.84	0.82	0.83	0.83	0.73	0.83	0.82	0.64
	precision	0.74	0.70	0.64	0.75	0.76	0.62	0.76	0.76	0.63
	Accuracy	0.75	0.71	0.64	0.75	0.71	0.60	0.76	0.76	0.59
	F-measure	0.79	0.76	0.72	0.79	0.76	0.67	0.79	0.74	0.63
	Stdev	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05
ISRI stemming and stop words	Recall	0.82	0.83	0.84	0.85	0.84	0.80	0.82	0.82	0.60
	precision	0.77	0.77	0.70	0.77	0.74	0.64	0.76	0.76	0.63
	Accuracy	0.77	0.73	0.71	0.78	0.75	0.64	0.76	0.76	0.59
	F-measure	0.80	0.80	0.76	0.81	0.79	0.71	0.79	0.74	0.62
	Stdev	0.03	0.04	0.07	0.04	0.06	0.07	0.02	0.02	0.08
Light stemming	Recall	0.78	0.78	0.64	0.79	0.81	0.77	0.80	0.79	0.74
	precision	0.70	0.70	0.63	0.73	0.73	0.62	0.73	0.73	0.59
	Accuracy	0.77	0.69	0.59	0.72	0.73	0.61	0.73	0.73	0.57
	F-measure	0.74	0.74	0.63	0.76	0.77	0.69	0.77	0.77	0.66
	Stdev	0.05	0.04	0.04	0.06	0.07	0.07	0.03	0.04	0.06
Light stemming and stop words	Recall	0.80	0.84	0.80	0.79	0.83	0.82	0.79	0.82	0.78
	precision	0.74	0.76	0.64	0.74	0.76	0.64	0.74	0.76	0.70
	Accuracy	0.73	0.77	0.64	0.73	0.71	0.64	0.73	0.76	0.77
	F-measure	0.77	0.80	0.71	0.76	0.76	0.72	0.76	0.74	0.74
	Stdev	0.07	0.07	0.06	0.07	0.06	0.07	0.07	0.07	0.07

4.5 Results based on sentiment words of presence and frequency features

This section presents the comparison results between different classifiers when we used Sentiment words of presence and frequency features (presence of violence word, the total frequency of violence words, presence of normal words, total frequency of normal words, ratio of the presence of violence to normal word and ratio of the total frequency of violence to normal words) that extracted from Arabic lexicon. In this case, there is no need to compare between preprocess files. Table 10 lists the results Based on Sentiment words of presence and frequency features.

Table 10. Results based on sentiment words of presence and frequency features.

	SVM	NB	KNN
Recall	0.75	0.68	0.73
Precision	0.69	0.76	0.62
Accuracy	0.68	0.70	0.60
F-measure	0.72	0.72	0.67
Stdev	0.06	0.06	0.05

Table 10 reveals that the best measured results recorded by SVM, followed by KNN and the worst performance is reported by NB. Moreover, classifies show stable models and the differences are insignificant.

4.6 Results based on Part of Speech (POS) features

This section represents the comparison results between different classifiers when we used part of speech features (total number of violence adjective, total number of violence verb, total number of violence noun, total number of normal adjective, total number of normal verb, total number of normal noun, the ratio of the total frequency of violence adjective to normal words, the ratio of the total frequency of violence verb to normal words, the ratio of the total frequency of violence noun to normal words) that extracted from Arabic lexicon so there is no need to compare between preprocess files. Table 11 lists the results Based on POS features. NB followed by SVM and finally KNN records the best results in terms of accuracy, precision, recall and F-measure.

Table 11. Results based on POS features.

	SVM	NB	KNN
Recall	0.72	0.76	0.66
Precision	0.71	0.75	0.77
Accuracy	0.72	0.72	0.70
F-measure	0.72	0.75	0.71
Stdev	0.04	0.06	0.05

4.7 Results based on other features

This section represents the comparison results between different classifiers when we used other features (presence of negation word in sentence, presence of violence emoticons and presence of normal emoticons in sentence). Table 12 depicts the results based on other features. It shows that SVM performs better than NB and KNN and produces more stable models.

Table 12. Results based on other features.

	SVM	NB	KNN
Recall	0.68	0.60	0.64
Precision	0.76	0.63	0.63
Accuracy	0.70	0.59	0.59
F-measure	0.72	0.62	0.63
Stdev	0.03	0.04	0.04

4.8 Results based on all features without n-gram

This section represents the comparison results between different classifiers when we used all features except n-gram feature. Table 13 represents the results Based on all features except n-gram.

Table 13. Results Based on all features except n-gram.

	SVM	NB	KNN
Recall	0.73	0.74	0.66
Precision	0.62	0.69	0.77
Accuracy	0.66	0.67	0.70
F-measure	0.67	0.71	0.71
Stdev	0.06	0.04	0.05

As shown in table 13, it reveals the best measured results when using NB classifier.

4.9 Results based on all features with n-gram feature on various preprocessing files

This section represents the comparison results between different classifiers when we used all features with the n-gram feature on various preprocessing techniques as presented in the Table 14.

Table 14. Results based on all features with n-gram feature on various preprocessing file

File name	Measures	Unigram+all features			Bigram+all features			Trigram+all features		
		SVM	***NB***	***KNN***	***SVM***	***NB***	***KNN***	***SVM***	***NB***	***KNN***
Normalization	Recall	0.81	0.85	0.66	0.81	0.85	0.65	0.81	0.85	0.66
	precision	0.76	0.75	0.76	0.76	0.75	0.77	0.76	0.75	0.77
	Accuracy	0.75	0.74	0.69	0.75	0.76	0.69	0.75	0.76	0.70
	F-measure	0.78	0.80	0.70	0.78	0.80	0.70	0.78	0.80	0.71
	Stdev	0.06	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.05
Stop words	Recall	0.81	0.86	0.74	0.82	0.86	0.77	0.82	0.86	0.76
	precision	0.75	0.75	0.75	0.76	0.75	0.73	0.75	0.75	0.74
	Accuracy	0.75	0.77	0.72	0.75	0.76	0.72	0.75	0.76	0.72
	F-measure	0.78	0.80	0.75	0.79	0.80	0.75	0.78	0.80	0.75
	Stdev	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.04	0.05
ISRI stemming	Recall	0.86	0.84	0.76	0.86	0.83	0.75	0.87	0.83	0.75
	precision	0.77	0.78	0.76	0.77	0.78	0.75	0.77	0.78	0.76
	Accuracy	0.78	0.81	0.73	0.78	0.77	0.72	0.78	0.77	0.73
	F-measure	0.81	0.78	0.76	0.81	0.80	0.75	0.81	0.80	0.76
	Stdev	0.03	0.04	0.04	0.04	0.05	0.05	0.04	0.05	0.05
ISRI stemming and stop words	Recall	0.87	0.85	0.75	0.85	0.82	0.76	0.85	0.82	0.76
	precision	0.78	0.78	0.76	0.78	0.76	0.75	0.78	0.76	0.75
	Accuracy	0.79	0.78	0.73	0.78	0.75	0.72	0.78	0.76	0.73
	F-measure	0.82	0.81	0.76	0.81	0.79	0.75	0.81	0.79	0.76
	Stdev	0.02	0.04	0.04	0.03	0.05	0.06	0.02	0.03	0.03
Light stemming	Recall	0.85	0.82	0.69	0.85	0.81	0.68	0.85	0.81	0.68
	precision	0.77	0.79	0.76	077	0.79	0.76	0.77	0.79	0.76
	Accuracy	0.78	0.78	0.70	0.78	0.77	0.70	0.78	0.78	0.70
	F-measure	0.81	0.81	0.72	0.81	0.80	0.72	0.81	0.80	0.72
	Stdev	0.06	0.05	0.05	0.03	0.05	0.05	0.03	0.04	0.06
Light stemming and stop words	Recall	0.85	0.83	0.78	0.85	0.82	0.69	0.85	0.83	0.66
	precision	0.77	0.77	0.72	0.77	0.77	0.76	0.77	0.77	0.77
	Accuracy	0.77	0.77	0.71	0.77	0.76	0.70	0.77	0.77	0.70
	F-measure	0.81	0.80	0.75	0.81	0.79	0.72	0.81	0.80	0.71
	Stdev	0.06	0.07	0.06	0.04	0.06	0.07	0.06	0.07	0.07

Table 14 shows that NB performs better than SVM and KNN on the normalization files regardless of the n-gram selected. It gives the same results using all features when n-gram is unigram, bigram and trigram. In the case of stop word file NB recorded better results too in all features with all n-gram variations. On the other hand, SVM is performing better than the NB and KNN in the other files (ISRI stemming, ISRI stemming with stop words, light stemming and light stemming with stop). We can notice that the maximum accuracy scored was by SVM classifier on ISRI stemming with stop words file with unigram and all features.

To sum up, in the experiment of N-gram feature with various preprocessing files, the optimized SVM outperformed other classification algorithms in ISRI stemming and stop word file out of six files when used with bigram feature. The experiment of using Sentiment words of presence and frequency features that extracted from lexicon on dataset shows that SVM outperformed other classification algorithms. Moreover,

when using POS features that are extracted from lexicon, NB outperformed the other classifiers. NB also outperformed other methods in the experiment that uses all feature extracted from lexicon with n-gram. Furthermore, the experiment of using all features that extracted from lexicon with N-gram feature on various preprocessing files shows that SVM outperformed the other methods in ISRI stemming and stop word file among the six files when used with unigrams feature. SVM again outperformed other methods when applied on ISRI stemming with stop words file with unigram and all features.

5 Conclusion and Future Work

This study uses sentiment analysis to detect violence over social media in Arabic language. Set of steps have been applied systematically to address this problem. First, data was collected from two of the popular social media web sites (Facebook and Twitter), then this data was annotated as (Violence or Normal). After collecting the data, set of preprocessing techniques applied to normalize the data. Then, ASL was adapted and modified to capture the objective of this study and features extraction process was conducted to distinguish the sentences. These sentences were classified using (SVM, NB and KNN). Finally, the model was evaluated using (recall, precision, accuracy, F-measure and study).

In the light of this methodology, research questions have been addressed in details, and it has been found that using sentiment analysis can significantly detect the violence in Arabic language (Jordanian dialect). also, it has been concluded that the using of the Arabic lexicon with modification would help to extract features for detecting violence, what distinguishes this lexicon is the ability to determine whether the sentences are normal or violence as well as it can determine POS in the words. Moreover, this research shows that combining features extracted from lexicon with features extracted from sentence such as n-gram generates models that are more accurate. In addition, the study discusses the effect of various preprocessing techniques on the performance of the generated models. The experiments proved that using SVM classifier on ISRI stemming with stop words file with unigram and all features gave best results compared with other experiments. In the future, the method can be extended to cover other Arabic dialects and the dataset can be expanded.

6 References

[1] Magu, Rijul, Kshitij Joshi, and Jiebo Luo. "Detecting the hate code on social media." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1. 2017.

[2] Biltawi, Mariam, Wael Etaiwi, Sara Tedmori, Amjad Hudaib, and Arafat Awajan. "Sentiment classification techniques for Arabic language: A survey." In 2016 7th International Conference on Information and Communication Systems (ICICS), pp. 339-346. IEEE, 2016. https://doi.org/10.1109/IACS.2016.7476075

[3] Nasukawa, Tetsuya, and Jeonghee Yi. "Sentiment analysis: Capturing favorability using natural language processing." In Proceedings of the 2nd international conference on Knowledge capture, pp. 70-77. 2003. https://doi.org/10.1145/945645.945658
[4] Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." Information Retrieval 2, no. 1-2 (2008): 1-135. https://doi.org/10.1561/1500000011
[5] Salem, F. "The Arab Social Media Report 2017: Social media and the Internet of Things: Towards data-driven policymaking in the Arab world." Dubai: MBR School of Government 7 (2017).
[6] Djuric, Nemanja, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. "Hate speech detection with comment embeddings." In Proceedings of the 24th international conference on world wide web, pp. 29-30. 2015. https://doi.org/10.1145/2740908.2742760
[7] Itani, Maher, Chris Roast, and Samir Al-Khayatt. "Corpora for sentiment analysis of Arabic text in social media." In 2017 8th international conference on information and communication systems (ICICS), pp. 64-69. IEEE, 2017. https://doi.org/10.1109/IACS.2017.7921947
[8] El-Makky, Nagwa, Khaled Nagi, Alaa El-Ebshihy, Esraa Apady, Omneya Hafez, Samar Mostafa, and Shimaa Ibrahim. "Sentiment analysis of colloquial Arabic tweets." In ASE BigData/SocialInformatics/PASSAT/BioMedCom 2014 Conference, Harvard University, pp. 1-9. 2014.
[9] Duwairi, Rehab, and Mahmoud El-Orfali. "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text." Journal of Information Science 40, no. 4 (2014): 501-513. https://doi.org/10.1177/0165551514534143
[10] Hammer, Hugo Lewi. "Detecting threats of violence in online discussions using bigrams of important words." In 2014 IEEE Joint Intelligence and Security Informatics Conference, pp. 319-319. IEEE, 2014. https://doi.org/10.1109/JISIC.2014.64
[11] Yadav, Shashank H., and Pratik M. Manwatkar. "An approach for offensive text detection and prevention in Social Networks." In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-4. IEEE, 2015. https://doi.org/10.1109/ICIIECS.2015.7193018
[12] Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. "A lexicon-based approach for hate speech detection." International Journal of Multimedia and Ubiquitous Engineering 10, no. 4 (2015): 215-230. https://doi.org/10.14257/ijmue.2015.10.4.21
[13] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In Proceedings of the NAACL student research workshop, pp. 88-93. 2016. https://doi.org/10.18653/v1/N16-2013
[14] Alhelbawy, Ayman, Poesio Massimo, and Udo Kruschwitz. "Towards a corpus of violence acts in arabic social media." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1627-1631. 2016.
[15] Mubarak, Hamdy, Kareem Darwish, and Walid Magdy. "Abusive language detection on Arabic social media." In Proceedings of the first workshop on abusive language online, pp. 52-56. 2017. https://doi.org/10.18653/v1/W17-3008
[16] Abdelfatah, Kareem E., Gabriel Terejanu, and Ayman A. Alhelbawy. "Unsupervised detection of violent content in arabic social media." Comput. Sci. Inf. Technol. (CS IT) (2017): 1-7. https://doi.org/10.5121/csit.2017.70401
[17] Abdul-Mageed, Muhammad, Mona Diab, and Sandra Kübler. "SAMAR: Subjectivity and sentiment analysis for Arabic social media." Computer Speech & Language 28, no. 1 (2014): 20-37. https://doi.org/10.1016/j.csl.2013.03.001
[18] Patro, S., and Kishore Kumar Sahu. "Normalization: A preprocessing stage." arXiv preprint arXiv:1503.06462 (2015).
[19] Al-Moslmi, Tareq, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. "Arabic senti-lexicon: Constructing publicly available language resources for Arabic

sentiment analysis." Journal of information science 44, no. 3 (2018): 345-362. https://doi.org/10.1177/0165551516683908

[20] Dillon, Martin. "Introduction to modern information retrieval: G. Salton and M. McGill. McGraw-Hill, New York (1983). xv+ 448 pp., $32.95 ISBN 0-07-054484-0." (1983): 402-403.

[21] Yong, Zhou, Li Youwen, and Xia Shixiong. "An improved KNN text classification algorithm based on clustering." Journal of computers 4, no. 3 (2009): 230-237. https://doi.org/10.4304/jcp.4.3.230-237

[22] Barbosa, Luciano, and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." In Coling 2010: Posters, pp. 36-44. 2010.

[23] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!" In Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, no. 1. 2011.

[24] Qaiser, Shahzad, Nooraini Yusoff, Farzana Kabir Ahmad, and Ramsha Ali. "Sentiment Analysis of Impact of Technology on Employment from Text on Twitter." International Journal of Interactive Mobile Technologies vol.14, no. 7 2020. https://doi.org/10.3991/ijim.v14i07.10600

[25] Scherr, Simon, Svenja Polst, Lisa Müller, Konstantin Holl, and Frank Elberzhager. "The perception of emojis for analyzing app feedback.", iJIM - Vol. 13, No. 2, 2019, (2019): 19-36. https://doi.org/10.3991/ijim.v13i02.8492

7 Authors

Monther Khalafat is a Jordanian computer scientist and the main contributor to this work, and was a student at the University of Jordan, King Abdullah II School for Information Technology, Department of Information Technology, Amman (Jordan). Email: monther_1987@hotmail.com

Dr. Ja'far Alqatawna is a Jordanian Dr. of Business Security at the University of Jordan, King Abdullah II School for Information Technology, Department of Information Technology, Amman (Jordan). Currently, he is on sabbatical leave at the Higher Colleges of Technology, Faculty of Computer Information Systems, Dubai, UAE. E-mails: j.alqatawna@ju.edu.jo, jalqatawna@hct.ac.ae

Prof. Rizik Al-Sayyed is a Jordanian Prof. of Networks, Databases, and Data Science at the University of Jordan, King Abdullah II School for Information Technology, Department of Information Technology, Amman (Jordan). E-mail: r.alsayyed@ju.edu.jo

Dr. Mohammad Eshtay is a Jordanian Dr. of Machine Learning at LTUC, from the University of Jordan, King Abdullah II School of Information Technology, Amman, Jordan. E-mail: m.eshtay@ltuc.com

Dr. Thaeer Kobbaey is a Jordanian Dr. of Data Mining and Big Data at The Higher Colleges of Technology, Dubai (UAE). E-mail: tkobbaey@hct.ac.ae

Article submitted 2021-04-01. Resubmitted 2021-05-17. Final acceptance 2021-05-17. Final version published as submitted by the authors.