Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-020-04447-0

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages

Mohammed A. AlGhamdi[1] · **Murtaza Ali Khan**[1]

Received: 11 September 2019 / Accepted: 27 February 2020

Abstract
With the widespread use of messaging via social networks such as Twitter, Instagram, and Facebook, it is becoming imperative for researchers to devise intelligent systems for data analytics in the range of domains like business, health, communication, security, etc. The complex morphological and syntactic structure of Arabic sentences makes them difficult to analyze. This paper presents an intelligent system to analyze Arabic tweets for detecting suspicious messages. We acquired Arabic tweet data from micro-blogging social network Twitter via Twitter Streaming Application Programming Interface and save it in a required file format. The system tokenizes and preprocesses the tweet dataset. Manual labeling is performed on tweet dataset for suspicious (label 1) and not-suspicious (label 0) classes. The labeled tweet dataset is used to train a classifier using supervised machine learning algorithms for the detection of suspicious activities. During the testing phase, the system processes unlabeled tweet data and detects either it belongs to a suspicious or not-suspicious class. We tested the system using six supervised machine learning algorithms: (1) decision tree, (2) *k*-nearest neighbors, (3) linear discriminant algorithm, (4) support vector machine, (5) artificial neural networks, and (6) long short-term memory networks. A comparative analysis in terms of accuracy, execution time, and confusion matrices of the six classifiers is presented. The execution speed of ANN is lowest. In terms of predicting correct results, the SVM performs best among all the classifiers and yields 86.72% mean accuracy. The major outcomes of this work are development of labeled dataset of Arabic tweets, an intelligent behavior analysis of tweets using six machine learning algorithms to detect suspicious messages, a comparative analysis of six machine learning algorithms, and a development of a statistical benchmark that can be used for future studies about the detection of crimes on social media.

Keywords Twitter · Arabic tweets · Social media · Machine learning · Intelligent systems · Supervised learning

1 Introduction

Social media is an integral element of modern life, and it is a global phenomenon. People all over the world use social media platforms to share knowledge and exchange ideas on almost any topic and domain of life. There are several social media platforms, such as Facebook, Instagram, and Reddit. We selected micro-blogging social network Twitter for our study because our data and work are related to Saudi Arabia, which has one of the highest percentages of Twitter users [1]. There are around 9.9 million Twitter users, and 77% of Internet users have a Twitter account [2]. Further, Twitter messages are short and to the point suitable for criminals to use it for their communication.

However, there are methods to identify suspicious messages on Twitter written in the English language [3,4]. But the syntax and semantics of the Arabic language are different from the English language, and the techniques that are suitable to the English language cannot be applied to the Arabic language directly. This study aims to develop a system to detect suspicious messages written in the Arabic language. Therefore, we acquired tweet data in the Arabic language. The proposed system can be used to monitor topical trends over time and then track specific communications based on predefined topics, keywords, profile types, and behavior to predict the potential of a user or a group of users for being involved in unwanted activities. In this work, we encountered

✉ Murtaza Ali Khan
makkhan@uqu.edu.sa

Mohammed A. AlGhamdi
maeghamdi@uqu.edu.sa

[1] College of Computer and Information Systems, Umm Al-Qura University, Mecca, Saudi Arabia

Published online: 10 March 2020

two major difficulties. The first is to extract the semantics of Arabic tweets written without diacritics. The second challenge was to preform stemming and lemmatization, i.e., to remove a suffix and prefix of a word, and reduces derived forms of a word to a common root word. The main contributions of this work are summarized as follows:

1. An organized and labeled dataset of Arabic tweets is developed. Tweets are collected, preprocessed, labeled, and stored in a ready to use format.
2. An intelligent behavior analysis of Arabic tweets is performed using six machine learning algorithms to detect suspicious messages.
3. A comparative analysis of six machine learning algorithms is presented. A system for real-life applications can be built based on our results and recommendations.
4. A statistical benchmark is developed that can be used for future studies about the detection of crimes on social media for Arabic language.

In the areas of learning and teaching, social media is utilized to enhance quality, increase reach, and facilitate interaction. The work of [5] integrates the qualitative analysis alongside data mining techniques by validating social media data sense-making. The study investigates the issues students encounter during the Twitter chat.

In the health care and medical sectors, social media is used to improve the services dispensed by health care facilities. In [6], a weighted-based semi-supervised learning technique is developed to identify adverse drug events (ADEs) from non-adverse drug events over social media network The study helps to impel pharmacovigilance for patients on social media. The work in [7] operates on the data obtained from social media (Twitter) to monitor the health of people over time through measuring the risk factors behavior and helping in advertising the planned health campaigns.

The influence of social media on the social life of people is not unusual. As an example, human behaviors during the momentous events are modeled in the work of [8]. In this work of Tyshchuk and Wallace, the conduct of people is studied to determine the elements of their behavior over social media, where numerical measurements are generated for those elements.

The Arab Spring of 2011 demonstrated that Twitter can play a momentous role in instigating and organizing disorder and violence. On the positive side, the information acquired from social media can be used to identify illegal activities, deter potential damage, and improve the safety of citizens. The security-related uses of such information include: (a) terrorism averting, (b) social disorder prevention, (c) control drug trafficking control, and (d) circumvention of sex offending and bullying problems.

It is evident from the above discussion that social scientists, researchers, educationalists, and people from other domains are studying to understand the usage, trends, and impact of social media on society and people. But most of the social media research is focused on personal and business needs, while relatively little work is reported on security and intelligence needs. This work contributes significantly to improve the security of the public by providing a frame to analyze Arabic tweets and detect suspicious messages.

The rest of the paper is organized as follows. A brief discussion about the Arabic language and its morphology is described in Sect. 2. Literature review is presented in Sect. 3. The architecture and methodology of our system are illustrated in Sect. 4. Terminologies and equations to measure the performance of the system are described in Sect. 5. Experiments and results are presented in Sect. 6. Section 7 analyzes results and gives an insight view of the proposed system. Section 9 draws the conclusions of our research and sets the directions for further work.

2 Arabic Language and Its Morphology

Arabic is the national language of a group of countries located in the Middle-East and North Africa. There are 22 countries grouped together and known as the Arab world. There are around 422 million people that speak the Arabic language. A recently conducted study [9] about the use of social media in the Arab region states that there are more than 2 million active Twitter users. The study also confirms that 72% of all tweets posted in the Arab region are written in Arabic. It is known that there are only twelve languages that are written from the right to the left including Arabic, Aramaic, Azeri, Divehi, Fula, Hebrew, Kurdish, N'ko, Persian, Rohingya, Syriac, and Urdu. Among these twelve languages, Arabic language has the largest share.

The alphabet of the Arabic language has 28 letters. Sentences of the Arabic language are divided into two different types; the first sentence is known as the nominal sentence (NS), while the second type is a verbal sentence (VS). In simple structure, the NS started with the subject and followed by a subject descriptor, while the VS begins with a verb, subject, or object. In the Arabic language, there are several diacritics which are usually placed below or above the letters and known in the Arabic language as harakat. In general, Arabic diacritics include Damma, Fatha, Kasra, Sukun, Shadda, Tanwin. Figure 1 shows three diacritics placed above or below the hypothetical letters shown as circles. Figure 2 illustrates how a diacritic is combined with an Arabic letter. In Arabic text, diacritics play a crucial role in the semantics of the language, as the meaning of the word depends on how the diacritics are applied [10].

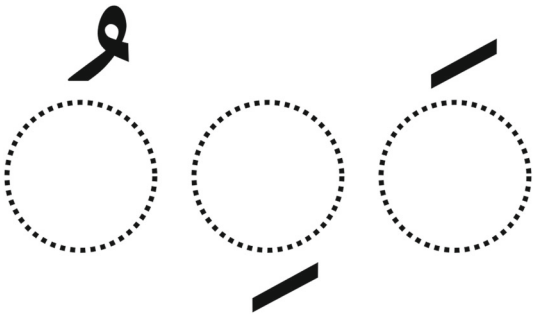


Fig. 1 Three Arabic diacritics are placed above or below the letters (circles)

Fig. 2 A diacritic is combined with a letter

However, the use of these diacritics is very limited when writing Arabic text in social media, which makes the task of tweet classification more difficult.

3 Literature Review

Twitter, an integral part of social media allows users to share text messages with their followers for personal, business, entertainment, and many other reasons. Despite its various benefits, there have been examples of social networks being misused for social unrest and illegal activities. In the literature, several authors mention the importance of Twitter. The authors of [11,12] reported that among social networks, Twitter receives larger attention because it is a platform that helps decision makers to know people's views and opinions related to a specific topic of interests. The authors of [11] highlight problems due to the short text of tweets and discuss the possibility of implementing successful solutions that can be used to overcome this problem. The authors of [13] emphasize and analyze the role of hashtags in tweets and suggest that a hashtag serves as both a tag of content and a symbol of membership of a community.

Use of social media for illegal activities and measures to control it is also the subject of the study of several authors. We briefly review a few such research work. In their paper, the authors of [14] argue that Internet provides new opportunities for all types of criminal activities including organized crime. Major criminal activities on social networks include (a) terrorism and extremism [15], (b) unrest and violence [16,17], (c) sex and drug trafficking [18], and (d) cyber-bullying and victimization [19]. Criminals exploit social networks for recruiting, organizing, and carrying out illegal activities. In the past, law enforcement agencies have relied upon Twitter and other social networks for intelligence gathering [20]. The authors of [21] emphasize the need for research in modeling to detect criminal activities on social networks. There is a great potential of social intelligence for the security and safety of communities and countries [22]. A bot also called robot is an automated computer program that runs over the Internet for a certain purpose. In Twitter, a bot can be used as a spammer to re-post messages for advertisements, spreading news, malicious websites, etc., [23]. A deep learning-based method to detect different types of abusive behavior such as racism, bullying, hate speech, sexism, and sarcasm is proposed by [24]. The system uses metadata and in conjunction with automatically extracted hidden patterns within the text of the tweets, to detect abusive behaviors. A method for automatically classifying hate speech on Twitter using deep learning is presented by [25]. The method combines convolutional neural networks (CNNs) and long short-term memory (LSTM) to detect hate speech. A deep learning framework for the classification of cyber-threats on Twitter is proposed by [26]. The framework is based on cascaded CNN architecture, a binary classifier, and a multi-class model. The system classifies of cyber-related tweets into multiple types of cyber-threats.

In the literature, several data mining and machine learning techniques are discussed to process textual data. These techniques can be used for various purposes such as the detection of suspicious activities, analysis of sentiments, and the marketing of products. A method of opinion mining on the Twitter social network is presented by [27]. The method converts the words' arrays to a numerical vector and applied a supervised learning approach to classify and analyze sentiments. [28] proposed a method that creates a ranked list of terms related to hashtags. Initially, the system learns users' perception of topic-term relationships, and later, when given a new tweet, it suggests suitable hashtags. General methods of Arabic text mining (e.g., sentiment analysis, topic modeling, classification, etc.) have been proposed and evaluated in the literature. These methods need to be evaluated for short text, and appropriate adaptations may be required for enhanced performance. A study of a sentiment-parsing algorithm of Arabic text (tweets, WhatsApp messages, etc.) on social networks is presented by [29]. A method of sentiment quantification for a given tweet is described by [30]. The method first determines the sentiment polarity (positive, negative, or neutral) and then identifies sentiments (happiness, love, sad, anger, etc.) within the tweet. To discover the sentiments of the Arabic phrases, the algorithm builds their associations with the words, based on word frequency and word co-location. Latent Dirichlet allocation (LDA) is a probabilistic model that can be used for topic modeling. The study of [31] uses LDA for Arabic topic identification. In this work, analysis of LDA is carried out at two levels: (1) stemming process and (2)

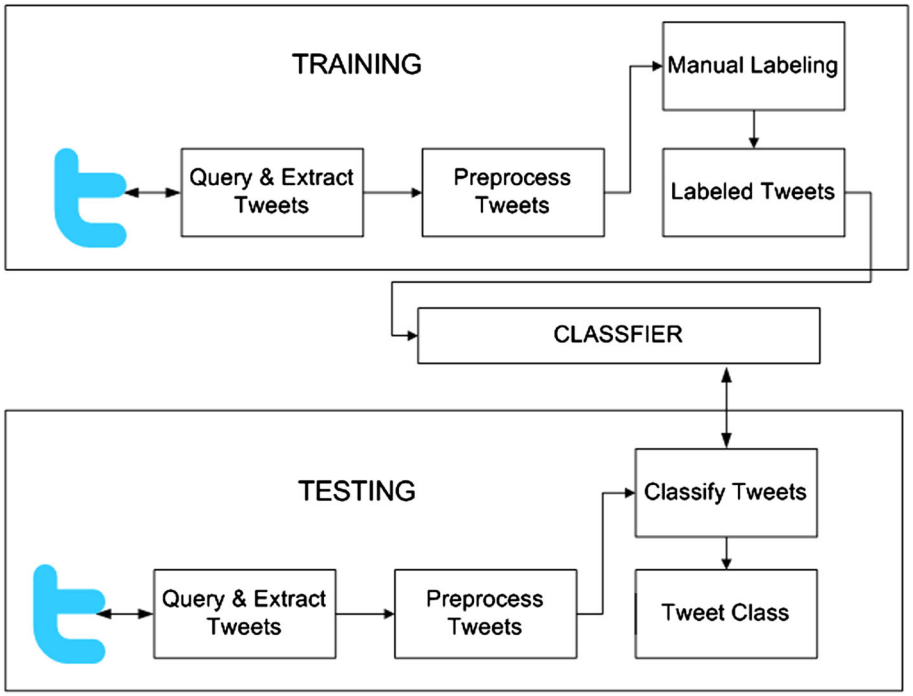


Fig. 3 System architecture

the choice of LDA hyper-parameters. A self-organizing map (SOM)-based clustering system for Arabic text documents that contain information about various categories of crimes is presented by [32]. Authors used a rule-based approach that exploits intransitive verbs and propositions to improve the clustering results. An LSTM–CNN-based deep learning method to summarize the text is proposed by [33]. An automated system to summarize Arabic text is proposed by [34]. The system extracts related segments from the text using the thematic structure of the document with the help of classifier and conceptual thesaurus. A rule-based method to identify Arabic named entity recognition (NER) in the crime domain is presented by [35]. A search tool to generate a report from the Arabic blogs for an input query is developed by [36]. An algorithm to aid Arabic information retrieval via query paraphrasing techniques is presented by [37]. One of these techniques employs the artificial bee colony algorithm, while the other utilizes a genetic algorithm. A study that compares deep recurrent neural network and support vector machine for sentiment analysis of Arabic hotels' reviews is presented by [38]. The authors report that the SVM outperforms the deep RNN.

Machine learning has been applied successfully to enhance the usage of big data in several different areas. The extreme learning machine (EELM) model is applied by [39] to investigate the stream and river flow prediction. The results of [39] conclude that the machine learning algorithm yields a more accurate predictive model for river flow. The work in [40] has employed the single artificial neural networks (ANNs), response surface methodology (RSM), and adaptive neuro-fuzzy inference system (ANFIS) to estimate and optimize the parameters related to the maximum biodiesel production yield (BPY). [41] proposed that an adaptive network-based fuzzy inference system (ANFIS) forecasts the discharges in Manwan hydroelectric dam in China. In their study, authors used data from 1953 to 1998 and from 1999 to 2003 for training and testing in monthly flow predictions. A survey in [42] shows that artificial neural networks can be employed successfully in providing an accurate forecasting model in the water quality areas. The ensemble empirical mode decomposition (EEMD) has been applied to predict time-series events such as runoff and rainfall in [43], and the results are effective. In the work of [44], solar radiation is estimated successfully by applying neural networks and fuzzy algorithms with focusing on several parameters such as the number of hidden layers in a neural network or determined weights. Table 1 summarizes the most significant works related to our research with the pros and cons of each approach.

4 System Architecture and Methodology

This section describes the architecture and methodology of the proposed system for acquiring, processing, training, classifying, and testing Arabic tweets for suspicious messages. Figure 3 shows the architecture diagram of the system. The system has three major modules: Training, Classifier, and Testing. The Training module queries Twitter using search terms/conditions and extracts the tweets. Tweets are preprocessed to remove unwanted data. Afterward, the manual labeling of tweets is performed. The labeled tweets are passed

Table 1 Table of related works with the pros and cons of each approach

References	Pros	Cons
[11,12]	Highlight problems due to the short text of tweets	The work lacks any comparative statistical analysis
[13]	Emphasize and analyze the role of hashtags in tweets	It does not analyze the role of text in tweets.
[14]	It is a broad discussion about the use of the Internet for organized crimes.	Twitter or other social media platforms are not discussed separately
[16,17]	Comprehensive studies about the role of Facebook and Twitter in political unrest	Discussion is limited only to political unrest in the Middle-East
[18,19,24,25]	In-depth analysis of one type of crime, e.g., drug, cyber-bullying, etc., on social media	Need to discuss crimes in general on social media
[29]	Sentiment parsing of Arabic language for capturing emotions from Arabic social media	Algorithm details are not given, and statistical results are not provided
[31]	Use Arabic stemming on LDA for topic identification	Only one classification algorithm (LDA) is used
[35]	Extracts named entities from Arabic crime documents	Rather than machine learning, it uses a fixed set of rules and has limited usage

to the Classifier module to train a machine learning algorithm (DT, KNN, LDA, SVM, ANN, and LSTM). The Testing module initially works like Training module, i.e., it queries the Twitter using search terms/conditions and extracts the tweets. Then, tweets are preprocessed to remove unwanted words/data. But afterward, the Testing module does not perform labeling. It classifies the unlabeled tweets using the pre-trained Classifier module. The following sections briefly describe the details of each module and the sub-module of the system.

4.1 Query and Extract Tweets

Twitter is a short messaging social network, and it provides Twitter Streaming API for developers, to use its services. The API allows a user to acquire (download) tweets data using *Access token* and *Access token secret*. Developers can search and download tweets related to the topic of interest using specified keywords in a language of choice (in our case Arabic). As our objective is to classify suspicious and not-suspicious tweets, we deliberately acquired a part of our dataset that composed of those tweets that come under the category of suspicious. The acquired data are in JSON (JavaScript Object Notation) format. We saved the tweet's data in Microsoft Excel using 8-bit Unicode Transformation Format (UTF-8) that supports the Arabic character set. Section 6 provides the details of the dataset.

4.2 Preprocess Tweets

4.2.1 Data Filtration

A tweet usually follows regular grammar, and it may contain some words or letters that should be neglected to improve the efficiency and quality of the proposed system. Data filtration removes unwanted characters or words from the tweets and reduces the training and testing time. We removed the following elements from the tweet dataset:

1. Punctuation marks, e.g., ; . , @ $? : !
2. Re-tweets
3. URL (uniform resource locator)
4. Media (images, videos, etc.)

4.2.2 Data Tokenization

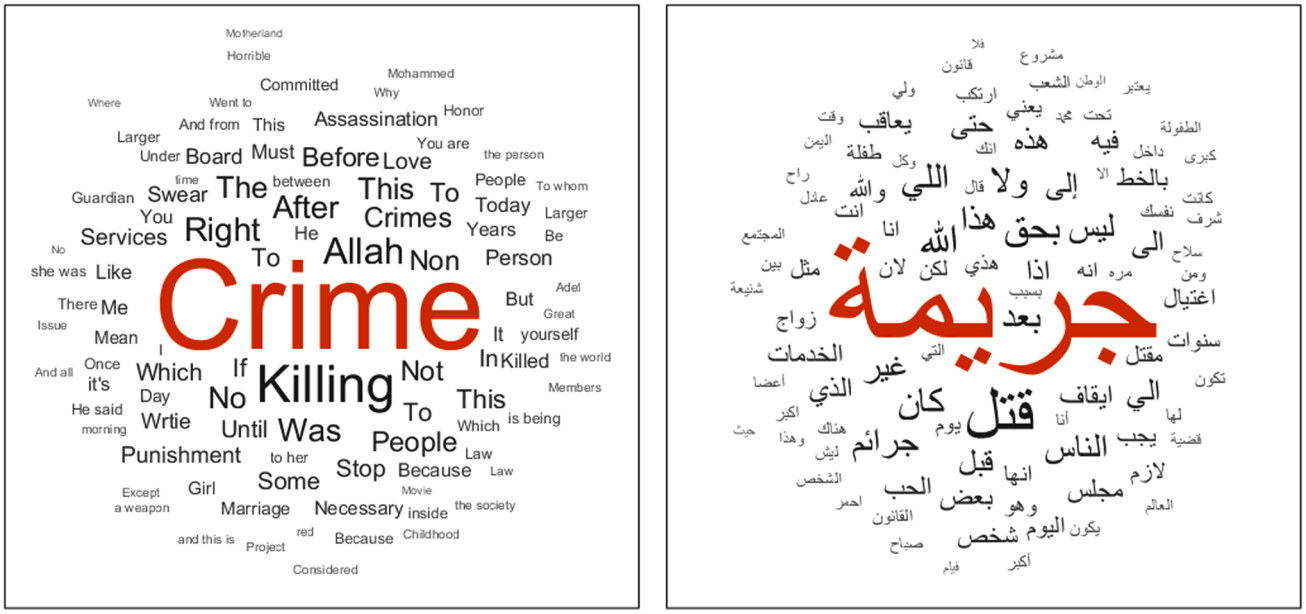
The Arabic language is written with spaces between its words and sentences. Tokens are individual words in a sentence. Tokens are separated by spaces. First, we tokenized the stream of tweet messages, then checked the length of each token, and removed tokens of length less than 3 characters, as their significance is very low, and mostly, they are conjunction (connecting) words. The tokenization process plays a crucial role in the stemming process.

4.2.3 Stemming and Lemmatization

Stemming and lemmatization, a natural language processing method, remove a suffix and prefix of a word and reduce derived forms of a word to a common base or root word. Messages, in our case Arabic tweets, use different forms of a word. We performed stemming and lemmatization to tweets before applying classification algorithms to the tweet dataset. Stemming and lemmatization may improve the accuracy of a classifier by reducing variations in words to their grammatical roots. Due to reduction in the size of vocabulary (bag of words), the indexing structure also becomes less complex, and consequently, accuracy of the classification system improves.

Table 2 Examples of Arabic tweets with translations and labels. Suspicious (label 1) and not-suspicious (label 0)

Arabic Tweet	English Translation	Label
جريمة يجب أن يعاقب عليها فاعلها	A crime's perpetrator must be punished	1 (suspicious)
عائلة في مدريد تتعرض للسطو في منزلهم من قبل مجرمين	A family in Madrid is being robbed in their home by criminals	1 (suspicious)
الموعد اللي ماتت أحلامنا فيه	The date when our dreams died	0 (not-suspicious)
أحيانا اللطف مع الناس جريمة ضد النفس	Sometimes kindness with people is a crime against the self	0 (not-suspicious)



(a) English translation **(b)** Arabic

Fig. 4 Word clouds of 100 most frequently occurring words

4.3 Manual Labeling

We performed manual labeling of tweets into two classes, suspicious (label 1) and not-suspicious (label 0). A tweet is labeled as suspicious if it needs further analysis by a security/intelligence analyst. This latter analysis includes looking further deep into the tweet and related data such as information about the person who tweeted, context, and time line. However, our study is limited only to label a tweet suspicious or not-suspicious. Fleiss's Kappa measure has been applied in this work with three experts' annotators [45]. A total of 72% of the tweets are categorized by the three experts with the same class, while two experts agree to classify 24% of the tweets. The remaining (4%) of the tweets are not agreed upon by the three annotators and are excluded from the data. Human judgment and intelligence play a factor in manual labeling, and it does not rely completely on a predefined set of words. To be consistent, we gave guidelines to the people performing labeling. On the downside, manual labeling is a tedious and long process and factors such as fatigue or personal opinion may affect it. In the future, we will automate this process with human verification. The study of [46] concludes that the performance of classification depends more on the size and quality of training data rather than the type of labeling, manual or automated. A few samples of Arabic tweets with their English translations and labels are shown in Table 2. Word clouds of 100 most frequently occurring words are shown in Fig. 4.

4.4 Machine Learning Models for Classification

There are several machine learning algorithms, supervised and non-supervised. For our work, supervised machine learning algorithms are suitable, as we can train the algorithm to identify suspicious messages. Supervised machine learning algorithms use a training dataset (label tweets) to construct a model that is used to classify the testing dataset (tweets without labels). Four of the methods (DT, KNN, LDA, and SVM) use the bag-of-words (BoW) model (also known as a term frequency counter) to represent tweets. The remaining two methods (ANN and LSTM) use word embedding where words or phrases from the tweets are mapped to vectors of

real numbers. The BoW models rely on the term frequency, i.e., the number of times a word occurs in a given text or document. Once the BoW is created, we trained a supervised classification model using the word frequency counts from the bag-of-words model and the labels.

4.4.1 Decision Tree (DT)

A decision tree (DT) is one of the most frequently used classification techniques [47] that takes a set of attribute values, i.e., a bag of words from tweets, as input and returns a decision, a single output value, i.e., a tweet is either suspicious or not-suspicious by applying a sequence of tests. Each internal node in the DT serves to test the input attributes values, W_i. The paths from the internal nodes are marked with the values of the input attributes, $W_i = v_{ik}$. Each leaf node in the DT designates the returned value, i.e., suspicious or not-suspicious.

4.4.2 *k*-Nearest Neighbors (KNN)

k-Nearest neighbor (KNN) is a supervised machine learning algorithm that can be used for classification. KNN works by leveraging similarities among examples (tweets). Having a set of training tweet data with labels, in our case, tweets with assigned labels of suspicious (1) and not-suspicious (0). When given a tweet without a label from testing data, it compares its vocabulary with those of trained tweets. Then, take the most similar tweets (the nearest neighbors) and look at the top K most similar tweets from the trained tweets. Take a majority vote from the K most similar tweets, and the majority is the new class assign to the tweet were asked to classify. In KNN, the value of K is usually less than 20.

4.4.3 Linear Discriminant Algorithm (LDA)

Fisher's linear discriminant algorithm (LDA) [48] is used to find a linear combination of features that separates two or more classes, in our case suspicious and not-suspicious tweets. Fisher proposed a maximizing function that yields a large separation between the projected class means (average), simultaneously gives a small variance within each class, and thus minimizes overlap of classes. LDA presumes that input classes produce data based on different Gaussian distributions. The LDA algorithm expresses one dependent variable as a linear combination of other features. LDA is used when classes (suspicious and not-suspicious) are known in advance. During the training of the classifier, LDA estimates the parameters of a Gaussian distribution for suspicious and not-suspicious tweets. During testing to determine the class of an unlabeled tweet, the classifier finds the class with the smallest mismatched cost.

4.4.4 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised machine learning algorithm that analyzes data (tweets). SVM is widely used for solving classification problems in the range of domains from textual data to images. An SVM training algorithm builds a model that assigns examples (text of tweets) into categories (suspicious and not-suspicious). The goal of SVM is to produce a tweet classification model, based on the training set of tweets, that predicts the testing set of tweets. In SVM, a nonlinear mapping of input tweets in a higher-dimensional feature space is done with the help of kernel functions. In feature space, a separation hyperplane is generated that is the solution to the classification of testing tweets. During the training phase of SVM, only a subset of the training dataset is needed and they are called support vectors. Sequential minimal optimization (SMO) algorithm [49] can be used for training the support vector machines.

4.4.5 Artificial Neural Networks (ANN)

Artificial neural network (ANN) is a biologically inspired machine learning model that resembles the operation of the human brain. In an ANN, many simple units work together in parallel with no centralized control unit. The weights between the units are the primary means of long-term information storage in neural networks. Updating the weights is the primary way the neural network learns new information. A simple ANN can be modeled by equation $Ax = b$. The A matrix is input tweet data in numeric form, and the b column vector represents labels, i.e., suspicious (1) and not-suspicious tweets (0) for each row in the A matrix. The weights on the neural network connections become x (the parameter vector). The behavior of ANN is controlled by its network architecture. A network's architecture can be defined (in part) by (1) number of neurons, (2) number of layers, and (3) types of connections between layers.

4.4.6 Long Short-Term Memory Networks (LSTM)

Long short-term memory (LSTM) networks are a variation in recurrent neural networks (RNNs). LSTM networks were introduced in 1997 by Hochreiter and Schmidhuber [50]. The critical component of the LSTM is the memory cell and the gates (input, output, and forget gates). The contents of the memory cell are modulated by the input gates and forget gates. Assuming that both of these gates are closed, the contents of the memory cell will remain unmodified between one time step and the next. The gating structure allows information to be retained across many time steps and consequently also allows gradients to flow across many time steps. This allows the LSTM model to overcome the vanishing gradient problem that occurs with most RNN models. We used LSTM

to identify suspicious and not-suspicious tweets due to their ability to classify character-level language models.

5 Terminologies

The following terminologies are used to measure the accuracy of our system.

- True positive (TP) are those tweets that correctly classified as suspicious.
- True negative (TN) are those tweets that correctly classified as not-suspicious.
- False positive (FP) are those tweets that incorrectly classified as suspicious.
- False negative (FN) are those tweets that incorrectly classified as not-suspicious.
- Accuracy is the ratio of tweets that are correctly classified to the total number of tweets classified, mathematically, Accuracy $= (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$.
- True positive rate (TPR) or sensitivity is the probability of correctly detecting a tweet as suspicious, mathematically, $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$.
- True negative rate (TNR) or specificity is the probability of correctly detecting a tweet as not-suspicious, mathematically, $\mathrm{TNR} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$.
- Positive predictive value (PPV) is the probability that a tweet label "suspicious" is actually suspicious, mathematically, $\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$.
- Negative predictive value (NPV) is the probability that a tweet label "not-suspicious" is actually not-suspicious, mathematically, $\mathrm{NPV} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FN})$.

6 Experiments and Results

We used MATLAB 2019a for our experiments on a 64-bit machine and Windows 10 operating system with 16 GB memory. Our tweet dataset consists of total 1555 tweets, out of which 826 tweets are suspicious, and 729 are not-suspicious. The accuracy of the classifier to detect suspicious and not-suspicious tweets is affected by several factors such as training and testing dataset, preprocessing, type of classifier, and the language of the text (Arabic in our case). We evaluated system accuracy using DT, KNN, LDA, SVM, ANN, and LSTM classifiers. Table 3 shows classification accuracy at various values of a holdout. The $P\%$ holdout means that the original tweet dataset is randomly partitioned into two sets containing $P\%$ of the tweets as the testing set and remaining $(1 - P)\%$ of tweets as the training set. For each classifier, we performed the simulations 100 times for each holdout value and then took the mean accuracy value. Among the 100 iterations, we computed the confusion matrices (CMs) of an arbitrary iteration for each classifier. Different classifiers may achieve maximum accuracy in a different iteration. Figure 5 shows accuracy and execution time of classifiers as line graphs at various values of holdout. Execution time includes the sum of partitioning, training, testing, and accuracy evaluation times in seconds. Since the execution time of LSTM is comparatively high compared to other classifiers, we have to show it in a separate subplot. Figure 6 shows confusion matrices (CMs) of six classifiers at 40% holdout value (i.e., 40% test data). Each CM consists of 3×3 cells. Following is the distribution of information in the cells of a CM.

Table 3 Mean accuracy of classification at various values of a holdout

Holdout%	DT	KNN	LDA	SVM	ANN	LSTM
20	85.53	73.92	77.23	86.72	80.96	74.34
30	85.36	75.38	76.99	86.09	80.64	72.66
40	84.59	73.44	76.65	85.61	81.54	74.53
50	84.07	72.92	76.40	85.52	80.87	72.74
60	84.75	71.51	75.61	85.32	79.74	72.95
70	84.68	70.98	76.39	85.21	78.46	69.04
80	84.43	69.31	77.62	85.39	75.76	71.95

- The top left four cells (2×2) of a CM show number of tweets and the percentage of the total number of tweets classified. In this 2×2 array of cells, rows correspond to the predicted class (determined by the classifier) and the columns correspond to the true class (as given in the known dataset of tweets). The diagonal cells correspond to tweets that are correctly classified (i.e., TP + TN). The off-diagonal cells correspond to incorrectly classified tweets (i.e., FP +FN).
- The rightmost column (1×3) of a CM shows the percentages of all the tweets "predicted" by the classifier.
- The bottom row (3×1) of a CM shows the percentages of all the tweets that "actually" belong to each class.
- The bottom right cell (1×1) of a CM shows the overall accuracy of the classifier.

7 Discussion

7.1 Analysis of Machine Learning Algorithms

From the experiments, it is evident that the accuracy of the SVM classifier is best, followed by DT, while the accuracy of LSTM and KNN is poor. In general, SVM is the first choice to classify natural language using the bag-of-words model. The textual data are high dimensional, and SVM performs mapping of the text data into high-dimensional feature spaces and

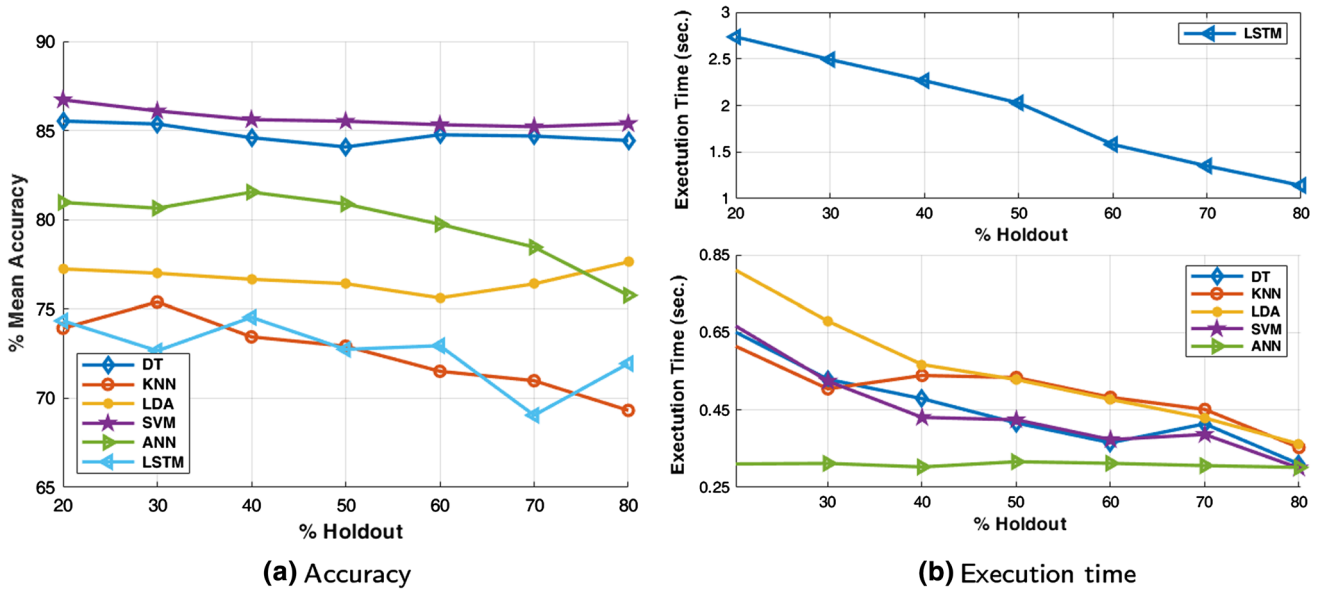


Fig. 5 Mean accuracy and mean execution time of classifiers as line graphs at various values of the holdout

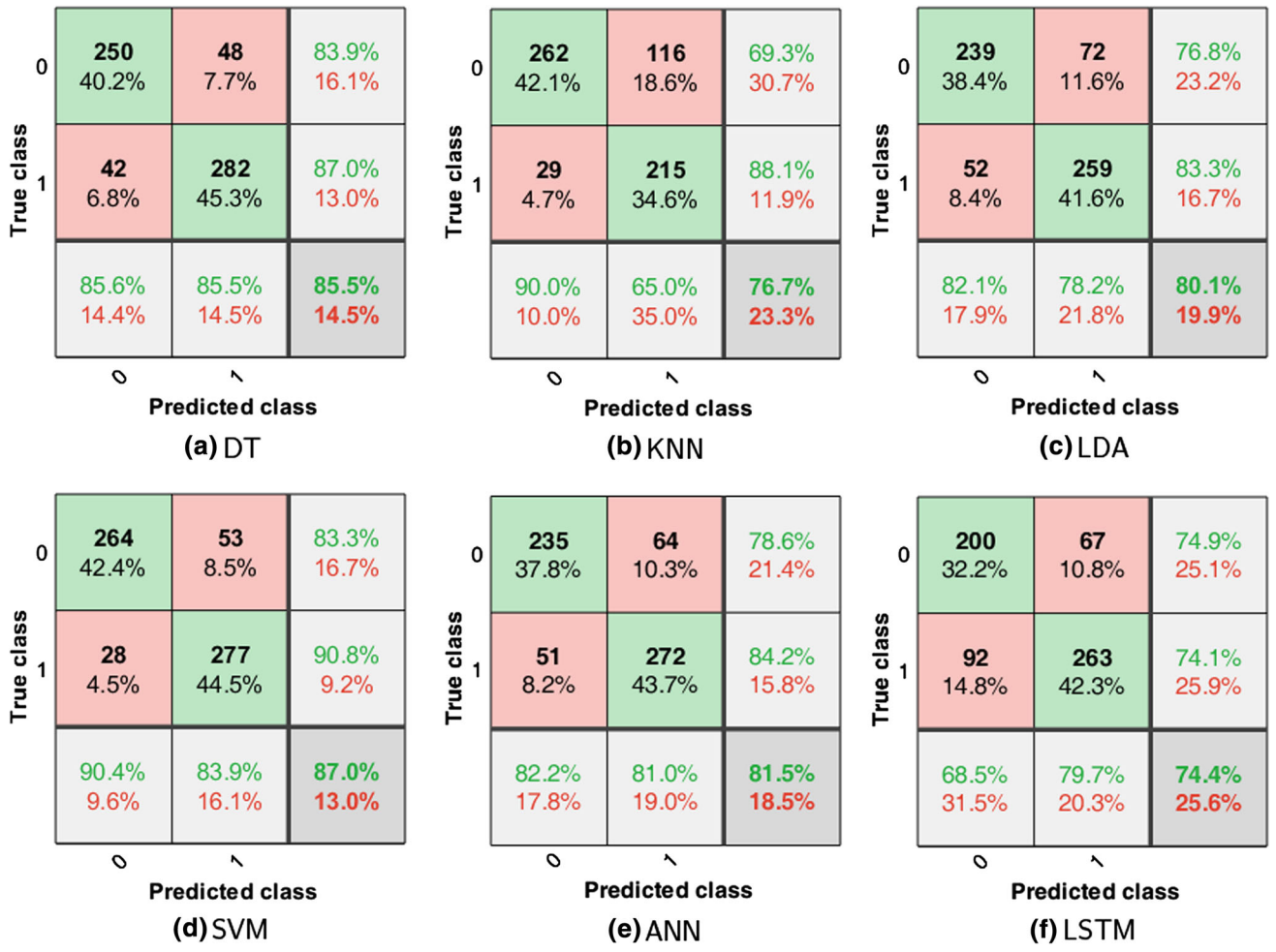


Fig. 6 Confusion matrices of test dataset at 40% holdout for any iteration (1–100) where the accuracy is maximum. Labels: 1 → suspicious, 0 → not-suspicious

Table 4 Statistics related to confusion matrices at 40% holdout (see Fig. 6)

Holdout%	DT%	KNN%	LDA%	SVM%	ANN%	LSTM%
TP	40.2	42.1	38.4	42.4	37.8	32.2
TN	45.3	34.6	41.6	44.5	43.7	42.3
FP	7.7	18.6	11.6	8.5	10.3	10.8
FN	6.8	4.7	8.4	4.5	8.2	14.8
TPR	85.6	90	82.1	90.4	82.2	68.5
TNR	85.5	65	78.2	83.9	81	79.7
PPV	83.9	69.3	76.8	83.3	78.6	74.9
NPV	87	88.1	83.3	90.8	84.2	74.4

Table 5 The standard deviation of accuracy in N iterations at various values of the holdout

Holdout%	DT	KNN	LDA	SVM	ANN	LSTM
20	1.85	1.69	1.97	1.59	3.00	2.15
30	1.22	1.28	2.35	0.90	1.56	2.91
40	1.50	2.60	2.58	1.07	1.44	1.59
50	0.90	1.16	1.31	0.80	2.03	3.65
60	1.28	1.77	1.94	1.08	2.55	3.29
70	1.04	1.64	2.21	0.41	1.77	4.34
80	1.56	2.37	2.41	0.77	2.03	2.79

then determines the decision boundary, i.e., where to draw the best hyperplane that divides the space into two subspaces, i.e., suspicious and not-suspicious. Other classification algorithms do not provide the optimal division of textual data and yield low accuracy compared to SVM. The accuracy of the decision tree to identify suspicious tweets is second only to SVM. DT is a good choice to classify data due to its simplicity (Yes/No) and the interpretation of results. But it has limited power to learn complicated rules that are needed to classify the textual data compared to SVM. The reason for the poor performance of KNN is that it computes the distance between words and chooses the class of closet words. But due to the inherent nature of the Arabic language, unrelated words can also be closed to each other, and KNN classifies them wrongly. LSTM a deep learning method based on a neural network is more suitable for big dataset that has a large number of classes. Our dataset is small, and it has only two classes, suspicious and not-suspicious.

7.2 Analysis of Confusion Matrices

The original dataset contains 826 suspicious tweets and 729 not-suspicious tweets. At 40% holdout value, suspicious and not-suspicious tweets are 330 and 292, respectively. The sum of all the accuracies for any classifier is 100. For example, for DT in Fig. 6a or Table 4, TP = 40.2%, TN = 45.3%, FP = 7.7%, and FN = 6.8%. Therefore, TP + TN + FP + FN = 40.2 + 45.3 + 7.7 + 6.8 = 100. To achieve overall high accuracy, it is desirable that true positive and true negative are higher and false positive and false negative are lower.

In Table 4, the first two values of column 4 are TP and TN of SVM, and their sum is highest (42.4 + 44.5 = 86.9) among all the classifiers. The third and fourth values in the column 4 of Table 4 are FP and FN of SVM. The sum of FP and FN is the lowest (8.5 + 4.5 = 13) for SVM among all the classifiers. In other words, SVM identifies tweets as suspicious and not-suspicious most accurately, and further, it does minimal mistakes in classifying suspicious tweets as not-suspicious and vice versa. This leads SVM to achieve the highest overall accuracy. DT performs quite close (but inferior) to SVM in all the factors. Consequently, the overall accuracy of DT is second only to SVM.

The TP rate of KNN is 42.1%, and it is second only to SVM, i.e., in most instances, KNN correctly classifies suspicious tweets as suspicious. However, the false positive rate of KNN is highest (18.6%) among all the classifiers, i.e., it incorrectly classifies a large number of not-suspicious tweets as suspicious. Therefore, the high FP rate of KNN causes its lowest overall accuracy.

7.3 Analysis of Standard Deviation (SD)

We performed the simulations 100 times for each holdout value and then took the mean accuracy value. The standard deviation at certain holdout value indicates the variations in the accuracy values from the mean accuracy at that holdout. The standard deviation at each holdout value is given in Table 5. It can be observed from Table 5 that the SD of SVM is lowest, while the SD of LSTM is highest. We can conclude from the SD values that the accuracy of SVM is most consistent within its 100 iterations. The SD of neural network-based methods, i.e., ANN and LSTM, is a bit higher compared to the other four methods.

7.4 Analysis of Execution Time

The mean execution time (Fig. 5b) of DT, KNN, LDA, SVM, and ANN does not exceed 0.85 s for any value of holdout. But the execution time of LSTM is comparatively high. A

large number of hidden units in LSTM cause its large execution time. ANN performs best with the lowest execution time due to its simple model of few neurons and two layers. In general, execution time decreases with the increase in the value of holdout. High holdout value means a large part of data is used for testing and a small part for training. In general, training takes more time than testing. For example, 20% holdout means 20% of the tweets are part of the testing set and 80% of tweets are part of the training set.

8 Limitation and Future Work

The limitation of the proposed work can be categorized into three different aspects; the first one is the use of relatively limited numbers of tweets, and in future, it is planned to increase them by digging more deeply into Twitter and extracting relevant data. The second aspect of limitation is the lack of further classification of tweets. In the future, we have the plan to broaden the system to categorize suspicious tweets into drug, sex, bullying, etc., classes. Finally, we used six machine learning algorithms in our study, and the system can be tested with ELMo [51], BERT [52], and other advanced models.

9 Conclusion

In this work, we presented a system that acquires, preprocesses, trains, classifies, and tests Arabic tweets to detect suspicious messages. Preprocessing involves data cleansing, filtration, stemming, and lemmatization. During the training stage, first manual labeling is performed on the tweet dataset to categorize it into suspicious and not-suspicious classes. Then, classifiers are trained using labeled tweets. During the classification stage, an unlabeled tweet is identified as suspicious or not-suspicious using the pre-trained classifier. We tested the system using six supervised machine learning algorithms, namely decision tree (DT), *k*-nearest neighbors (KNNs), linear discriminant algorithm (LDA), support vector machine (SVM), artificial neural networks (ANNs), and long short-term memory networks (LSTMs). We presented and analyzed the accuracy, execution time, and confusion matrices of the six classifiers. Simulation results suggest that the SVM classifier is the most favorable choice due to its higher accuracy and lower execution time compared to other classifiers. In a real-life scenario, the system shall be used in conjunction with final human judgment.

References

1. Bahkali, S.; Almaiman, A.; Bahkali, A.; Almaiman, S.; Househ, M.S.; Alsurimi, K.: The role of social media in promoting women's health education in saudi arabia. In: ICIMTH, pp. 259–262 (2015)
2. Albalawi, Y.; Sixsmith, J.: Identifying twitter influencer profiles for health promotion in Saudi Arabia. Health Prom. Int. **32**(3), 456–463 (2015). https://doi.org/10.1093/heapro/dav103
3. Dhavase, N.; Bagade, A.M.: Location identification for crime amp; disaster events by geoparsing twitter. In: International Conference for Convergence for Technology-2014, pp. 1–3 (2014). https://doi.org/10.1109/I2CT.2014.7092336
4. Wang, X.; Gerber, M.S.; Brown, D.E.: Automatic crime prediction using events extracted from twitter posts. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, pp. 231–238. Springer, Berlin (2012)
5. Chen, X.; Vorvoreanu, M.; Madhavan, K.: Mining social media data for understanding students' learning experiences. IEEE Trans. Learn. Technol. **7**(3), 246–259 (2014)
6. Liu, J.; Wang, G.; Chen, G.: Identifying adverse drug events from social media using an improved semi-supervised method. IEEE Intell. Syst. **34**, 66 (2019)
7. Sidana, S.; Amer-Yahia, S.; Clausel, M.; Rebai, M.; Mai, S.T.; Amini, M.R.: Health monitoring on social media over time. IEEE Trans. Knowl. Data Eng. **30**(8), 1467–1480 (2018)
8. Tyshchuk, Y.; Wallace, W.A.: Modeling human behavior on social media in response to significant events. IEEE Trans. Comput. Soc. Syst. **5**(2), 444–457 (2018). https://doi.org/10.1109/TCSS.2018.2815786
9. Salem, F.: Social media and the internet of things towards data-driven policymaking in the arab world: potential, limits and concerns. The Arab Social Media Report, Dubai: MBR School of Government, vol. 7 (2017)
10. Abed, S.; Alshayeji, M.; Sultan, S.: Diacritics effect on arabic speech recognition. Arab. J. Sci. Eng. (2019). https://doi.org/10.1007/s13369-019-04024-0
11. Kateb, F.; Kalita, J.: Classifying short text in social media: Twitter as case study. Int. J. Comput. Appl. **111**(9), 1–2 (2015). https://doi.org/10.5120/19563-1321
12. Maynard, D.; Bontcheva, K.: Challenges of evaluating sentiment analysis tools on social media. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1142–1148. LREC (2016)
13. Yang, L.; Sun, T.; Zhang, M.; Mei, Q.: We know what@ you# tag: does the dual role affect hashtag adoption? In: Proceedings of the 21st International Conference on World Wide Web, pp. 261–270. ACM (2012)
14. Lavorgna, A.: Organised crime goes online: realities and challenges. J. Money Laund. Control **18**(2), 153–168 (2015). https://doi.org/10.1108/jmlc-10-2014-0035
15. Taylor, R.W.; Fritsch, E.J.; Liederbach, J.: Digital Crime and Digital Terrorism. Prentice Hall Press, Upper Saddle River (2014)
16. Goodman, S.D.: Social media: The use of facebook and twitter to impact political unrest in the middle east through the power of collaboration. Technical Report, Faculty of the Journalism Department, California Polytechnic State University, San Luis Obispo, Senior Project (2011)
17. Bruns, A.; Highfield, T.; Burgess, J.: The Arab Spring and its social media audiences: English and Arabic Twitter users and their networks. In: Cyberactivism on the Participatory Web, pp. 96–128. Routledge, Abingdon (2014)
18. Coscia, M.; Rios, V.: How and where do criminals operate? Using Google to track Mexican drug trafficking organizations. In: CID Research Fellow and Graduate Student Working Paper (2012)
19. Dinakar, K.; Reichart, R.; Lieberman, H.: Modeling the detection of textual cyberbullying. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
20. Heverin, T.; Zach, L.: Twitter for city police department information sharing. Pro. Am. Soc. Inf. Sci. Technol. **47**(1), 1–7 (2010). https://doi.org/10.1002/meet.14504701277

21. Frank, R.; Cheng, C.; Pun, V.: Social Media sites: New Fora for Criminal, Communication, and Investigation Opportunities. Public Safety Canada, Ottawa (2011)
22. Wang, F.Y.; Carley, K.M.; Zeng, D.; Mao, W.: Social computing: From social informatics to social intelligence. IEEE Intell. Syst. **22**(2), 79–83 (2007). https://doi.org/10.1109/mis.2007.41
23. Chu, Z.; Gianvecchio, S.; Wang, H.; Jajodia, S.: Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Trans. Depend. Secure Comput. **9**(6), 811–824 (2012)
24. Founta, A.M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; Leontiadis, I.: A unified deep learning architecture for abuse detection. In: Proceedings of the 10th ACM Conference on Web Science, pp. 105–114. ACM (2019)
25. Zhang, Z.; Robinson, D.; Tepper, J.: Detecting hate speech on twitter using a convolution-gru based deep neural network. In: European Semantic Web Conference, pp. 745–760. Springer (2018)
26. Behzadan, V.; Aguirre, C.; Bose, A.; Hsu, W.: Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 5002–5007 (2018). https://doi.org/10.1109/BigData.2018.8622506
27. Sohrabi, M.K.; Hemmatian, F.: An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study. Multimed. Tools Appl. pp. 1–20 (2019)
28. Tariq, A.; Karim, A.; Gomez, F.; Foroosh, H.: Exploiting topical perceptions over multi-lingual text for hashtag suggestion on twitter. In: The Twenty-Sixth International FLAIRS Conference (2013)
29. Bani-Hani, A.; Majdalawieh, M.; Obeidat, F.: The creation of an arabic emotion ontology based on e-motive. Proc. Comput. Sci. **109**, 1053–1059 (2017). https://doi.org/10.1016/j.procs.2017.05.383
30. Bouazizi, M.; Ohtsuki, T.: Multi-class sentiment analysis in twitter: what if classification is not the answer. IEEE Access **6**, 64486–64502 (2018)
31. Naili, M.; Chaibi, A.; Ghézala, H.: Arabic topic identification based on empirical studies of topic models. Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées **27** (2017)
32. Alruily, M.; Ayesh, A.; Al-Marghilani, A.: Using self organizing map to cluster arabic crime documents. In: Proceedings of the International Multiconference on Computer Science and Information Technology. IEEE (2010). https://doi.org/10.1109/imcsit.2010.5679616
33. Song, S.; Huang, H.; Ruan, T.: Abstractive text summarization using LSTM-CNN based deep learning. Multimed. Tools Appl. **78**(1), 857–875 (2018). https://doi.org/10.1007/s11042-018-5749-3
34. Hammo, B.H.; Abu-Salem, H.; Evens, M.W.: A hybrid arabic text summarization technique based on text structure and topic identification. Inte. J. Comput. Process. Lang. **23**(01), 39–65 (2011). https://doi.org/10.1142/s1793840611002206
35. Asharef, M.; Omar, N.; Albared, M.: Arabic named entity recognition in crime documents. J. Theor. Appl. Inf. Technol. **44**(1), 1–6 (2012)
36. Walid, M.; Ali, A.; Darwish, K.: A summarization tool for time-sensitive social media. In: Proceedings of the international multiconference on computer science and information technology, pp. 2695–2697. Proceedings of the 21st ACM international conference on Information and knowledge management (2012)
37. Al-Dayel, A.; Ykhlef, M.: Enhanced arabic document retrieval using optimized query paraphrasing. Arab. J. Sci. Eng. **40**(11), 3211–3232 (2015). https://doi.org/10.1007/s13369-015-1797-4
38. Al-Smadi, M.; Qawasmeh, O.; Al-Ayyoub, M.; Jararweh, Y.; Gupta, B.: Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. J. Comput. Sci. **27**, 386–393 (2018). https://doi.org/10.1016/j.jocs.2017.11.006
39. Yaseen, Z.; Sulaiman, S.; Deo, R.; Chau, K.W.: An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol. **569**, 387–408 (2019)
40. Najafi, B.; Ardabili, S.; Shamshirband, S.; Rabczuk, T.: Application of anns, ANFIS and RSM for estimating and optimizing parameters affect the biodiesel production yield and cost. Eng. Appl. Comput. Fluid Mech. **12**(1), 611–624 (2018)
41. Cheng, C.T.; Lin, J.Y.; Sun, Y.G.; Chau, K.: Long-term prediction of discharges in Manwan hydropower using adaptive-network-based fuzzy inference systems models. In: Lecture Notes in Computer Science, pp. 1152–1161. Springer, Berlin, Heidelberg (2005). https://doi.org/10.1007/11539902_145
42. Fotovatikhah, F.; Herrera, M.; Shamshirband, S.; ardabili, S.; Jalil Piran, M.: Survey of computational intelligence as basis to big flood management: challenges, research directions and future work. Eng. Appl. Comput. Fluid Mech. **12**(1), 411–437 (2018)
43. Chuan Wang, W.; Wing Chau, K.; Qiu, L.; Bo Chen, Y.: Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. Environ Res Hydrol Water Resour **139**, 46–54 (2015)
44. Moazenzadeh, R.; Mohammadi, B.; Shamshirband, S.: Coupling a firefly algorithm with support vector regression to predict evaporation in Northern Iran. Eng. Appl. Comput. Fluid Mech. **12**, 584–597 (2018)
45. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378 (1971)
46. Mozetič, I.; Grčar, M.; Smailović, J.: Multilingual twitter sentiment classification: the role of human annotators. PLoS ONE **11**(5), e0155036 (2016). https://doi.org/10.1371/journal.pone.0155036
47. Russell, S.J.; Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson Education Limited, Malaysia (2016)
48. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
49. Platt, J.C.: Using analytic QP and sparseness to speed training of support vector machines. In: Advances in Neural Information Processing Systems, pp. 557–563 (1999)
50. Hochreiter, S.; Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
51. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the NAACL, pp. 2227–2237. Association for Computational Linguistics (2018)
52. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint arXiv:1810.04805