Evolutionary Intelligence (2021) 14:811–822
https://doi.org/10.1007/s12065-020-00458-w

SPECIAL ISSUE

Machine learning approach for threat detection on social media posts containing Arabic text

Shatha AbdulAziz AlAjlan[1] · **Abdul Khader Jilani Saudagar**[1]

Received: 2 January 2020 / Revised: 17 June 2020 / Accepted: 13 July 2020 / Published online: 4 August 2020

Abstract
Recently, social media has become a part of daily people's routine. People frequently share images, text, and videos in social media (e.g., Twitter, Snapchat, Facebook, and Instagram). Consequently, there is a demand for an automated method to monitor and analyze the shared social media content. This research developed a method that aims to detect any threat in the images or comments in the shared content. Instagram has gained popularity as the most famous social media website and mobile application for media sharing. Instagram enables users to upload, view, share, and comment on a media post (image or video). There are many unwanted contents in Instagram posts, such as threats, which may cause problems for society and national security. The purpose of this research is to construct a model that can be utilized to classify Instagram content (images and Arabic comments) for threat detection. The model was built using Convolutional Neural Network, which is a deep learning algorithm. The dataset was collected utilizing the Instagram API and search engine and then labeled manually. The model used was retrained on the images and comments training set with the classes of threat and non-threat. The results show that the accuracy of the developed model is 96% for image classification and 99% for comment classification. The result of this research will be useful in tracking and monitoring social media posts for threat detection.

Keywords Detection · CNN · Images · Comments · Threat · Transfer learning · TensorFlow · Inception v-3

1 Introduction

Social media has changed the mode of communication between people. Most people frequently share text, images, and videos in social media (e.g., Twitter, Instagram, Snapchat and Facebook). Therefore, utilizing social media as a method of communication between people has become a part of their daily routine. Shared content in social media posts needs to be tracked to determine what they contain through post information extraction. Consequently, there is a demand for automated processes and methods to monitor and analyze social media content. The researchers in [39] presented the Extracting the Meaning of Terse Information in a Geo-Visualization of Emotion (EMOTIVE) system, which is a project funded by the Defense Science and Technology Laboratory (DSTL), as a tool for national security and monitoring tasks to analyze emotional public responses to provide assistance regarding automation of the inefficient current monitoring tasks on Twitter. The researcher in [23] presented a tool to monitor and analyze social media, especially Twitter, which they named the Screening of New Media (SNM) tool. This tool helps the authorities understand citizens' reactions to alert messages about crises by utilizing sentiment analysis (SA) techniques for classifying emotional states. After reviewing the latest research, it is evident there is a lack of research on tracking shared social media posts that contain text and images, especially Arabic text.

Every day, billions of images and comments are uploaded and shared on social media. Instagram is a popular social media website and mobile application that is widely used for image and video sharing. It gives users an immediate way to capture and share images and videos of their life moments and add comments under those media. The number of users of Instagram is around one billion, and they are "Instagramming" more than 95 million posts per day [19]. Many people in Arabic-speaking countries utilize Instagram

✉ Abdul Khader Jilani Saudagar
aksaudagar@imamu.edu.sa

Shatha AbdulAziz AlAjlan
samalajlan@imamu.edu.sa

[1] Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

frequently, which makes it suitable for this research. Due in part to its nature, Instagram has become an important means of communicating threat activity. The threat in Instagram posts has become one of the sources of threats that threaten any country's national security or teenagers or children at home. This research aims to detect the threat in the Instagram images and comments. To achieve this, deep learning techniques must be utilized. Generally, social media images have received little attention, and therefore there is a need to classify them. Consequently, a classifier will be developed that will assist in detecting the threat in the Instagram posts (images and text comments) automatically.

This research will help determine whether text and images shared between users in Instagram are relevant to the discussion between parties or not. To the best of the researchers' knowledge, some threat detection techniques are utilized currently to analyze social media posts in English, but no efficient technique has been proposed for analyzing and monitoring Arabic content on social media (images, text, and videos). Therefore, this research will propose an approach to extract image content and classify it into threat and non-threat categories. Moreover, this research will detect any Arabic threat-related keywords in the comments. The results of this research will be useful in tracking and monitoring social media posts that contain images and comments in Arabic for threat detection.

The paper is organized as follows. Section 2 discusses background work related to the current state of social media monitoring and threat detection, feature extraction and feature classification, image classification methods, and deep learning approaches. Section 3 outlines and illustrates the methodology and experiment. The results are presented in Sect. 4, and the conclusion and suggestions for future work are provided in Sect. 5.

2 Background

2.1 Current state of social media tracking and threat detection

Natural Language Processing (NLP) and Sentiment Analysis (SA) tools were utilized in most of the previous studies on tracking and threat detection that focused on the social media text content. The researcher in [23] presented a tool that they referred to as SNM. The SNM tool is utilized for analyzing and tracking social media content in crises for the government authorities to analyze citizens' responses to alert messages. To construct this tool, the researchers needed to utilize APIs and web scraping methods to retrieve the relevant data. SA was then utilized for classifying posts' emotions. Twitter and Facebook were the primary social media websites considered in this study. However, the researchers in [24] proposed a method to detect users' behavior toward the law and the Greek authorities in YouTube comments. Two categories of comments were classified based on words and phrases related to various Greek jargon terms: positive for neutral behavior and negative for reluctant behavior. The researchers utilized a machine learning algorithm, logistic regression (LR), to carry out the classification.

The researchers in [39] presented the EMOTIVE system, a project funded by the DSTL. The EMOTIVE system is a tool utilized to carry out national security and monitoring tasks. It is used to analyze emotional public responses to assist in automating the ineffective text-based monitoring tasks in Twitter. The steps in this system are as follows. First, by checking the Twitter trending topics, the keyword/key phrase is filtered and extracted. Next, the geolocation and the tweets' emotion and tone need to be detected. Finally, a visualization of the user interface needs to be created.

The researcher in [32] aimed to gain an understanding of the social media power in the decision-making in the commercial industry to improve the quality of products and services and gain competitive advantage. The author's work focuses on the usage of text mining in the pizza sector of Twitter and Facebook to study the loyalty between customers and three big pizza restaurants. The findings of his research showed that social media plays a significant role in ensuring customer loyalty and satisfaction. Although the above studies on social media monitoring and analysis achieved good results, they did not explore threat detection based on images and text content together.

2.2 Feature extraction and feature classification

In the area of image processing, features play a vital role. Image classification involves classifying images into categories [41]. This study shows that there are specific steps when classifying images, as shown below in Fig. 1. The first step is preprocessing, which is applied before any image analysis processes are performed. In this step, image normalization is conducted, histogram equalization is carried out, and noise filtering and segmenting are performed. The next two steps, feature extraction and feature classification, are briefly

Fig. 1 Image classification process

explained in the next subsection. The basic step for classification is feature extraction, which involves extracting a set of features. The feature extraction process aims to extract the most important information from an image [28]. This important information is known as a feature, and it will be retained for the classification step [25].

2.2.1 Feature extraction

While classifying images, extracting the features is an essential step. In this step, the relevant data from the image are extracted and highlighted. Those relevant features form feature vectors that are fed into a suitable classifier. Features can include color, texture, and shape [42]. The major goal of feature extraction is maximizing the recognition rate with minimum elements [28]. Some of the widely used feature extraction techniques are described below.

2.2.1.1 Histogram oriented gradient (HOG) filter This technique focuses more on shape and works as encoded edges of an image in conjunction with their directions. For example, it extracts the edges of the facial features in an image, such as brows, mouth, and nose. The HOG filter is applied after dividing the image into smaller sub-images. Then, within each sub-image, a histogram for gradient direction is compiled. Finally, one vector of histograms becomes the final HOG feature [7].

2.2.1.2 Local binary pattern (LBP) This technique focuses more on image texture analysis. The image in this technique is split into small areas that the LBP histograms extracted into a single feature histogram. To extract the LBP features rapidly, a single scan through the image is carried out [14].

2.2.1.3 Haar-like features The Haar wavelet is a strong image feature descriptor that is used for object detection. The 2D Haar is decomposed of a square image with pixels of n2 size of. This square contains n2 wavelet coefficients, with each one corresponding to a separate wavelet. One wavelet is the whole image mean value of pixel intensity; the rest of the wavelets are computed by taking the difference in mean values of intensity for all adjacent squares horizontally, diagonally, or vertically. Every wavelet is restricted by the (x, y) wavelet location and the wavelet width and wavelet height to be aligned on a power of 2 [44].

2.2.2 Feature classification

After feature extraction, the main phase in identifying the image class is feature classification. It is necessary for the extracted features to be categorized into a proper class. The input into this step is the output of the feature extraction step.

2.2.2.1 Support vector machine (SVM) In the machine learning discipline, SVM is among the most common classifiers. It is usually more efficient than other classifiers. In [30], the authors explain that "Machine learning algorithms receive input data during a training phase, build a model of the input and output a hypothesis function that can be used to predict future data." Figure 2 illustrates the structure of the SVM model utilized with a simple classification problem that has two classes.

2.3 Hand-crafted features image classification methods

2.3.1 Viola–Jones algorithm

The researchers in [43] invented an algorithm for face detection. This detection technique is based on the Haar-like features that are determined by scale, position, and shape. First, for fast feature evaluation, integral image, which is a new representation of the image, is introduced to allow fast feature extraction on many scales. Second, the most important features are selected from the large Haar-like features. This algorithm does not work on all face images, however, as it only detects the frontal upright faces, as illustrated in Fig. 3.

2.3.2 Utilizing histogram oriented gradient (HOG)

Similar to the Viola–Jones algorithm, utilizing hand-coded features, the researchers in [7] studied the feature sets for human detection and found that HOG feature descriptors performed more efficiently than other feature descriptors available at that time. HOG works as follows. For every single pixel, it looks at every surrounding pixel to check how dark the current pixel is by comparing it with the surrounding pixels. Then, this pixel is replaced by an arrow called a gradient, which points to the direction of darker pixels

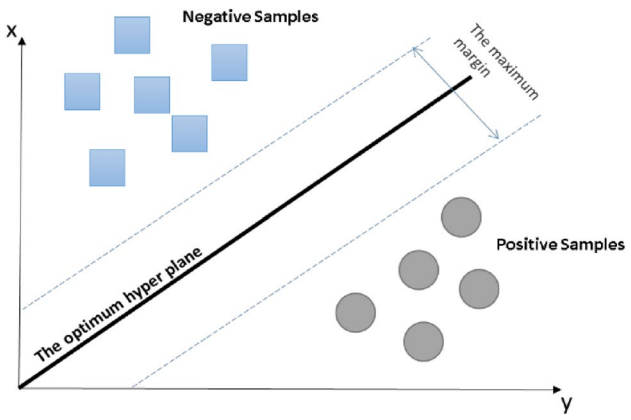


Fig. 2 SVM classifier for two classes (positive and negative). (*Source*: [12])

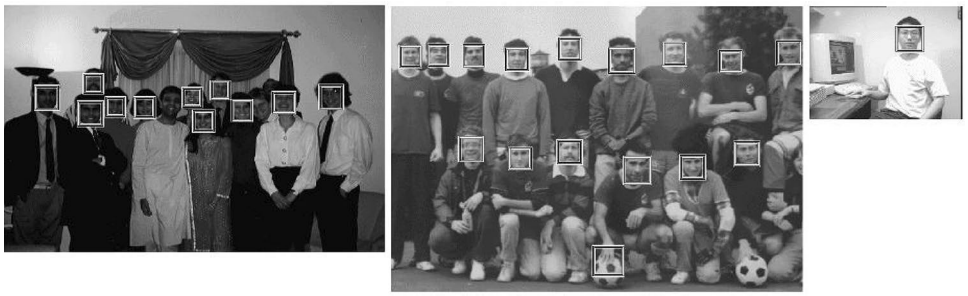


Fig. 3 Output of Viola–Jones face detector. (*Source*: [43])

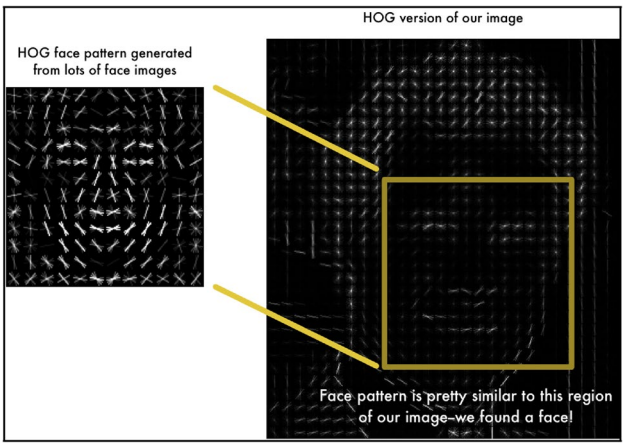


Fig. 4 HOG face detection. (*Source*: [7])

starting from the current pixel. It repeats this process for each pixel. Subsequently, in the entire image, the gradients show the flow from light to dark. Finally, the original image is converted into a simple representation that looks like the structure of a face (see Fig. 4). A face is detected by comparing this result with a known HOG pattern that was extracted from other training faces using the SVM classifier.

One problem with these traditional hand-crafted methods is that the feature extraction is separate and totally detached from the feature classification. This means that the classification is poor and suffers significantly if the chosen features cannot be those that are needed to distinguish between categories. Another problem with traditional hand-crafted methods is that they do not work in the same way as humans learn to recognize things. While a child is growing up, they process data and learn to identify things. This is the idea behind deep learning, though there are no hand-crafted features. The feature extraction and classification are combined to work as one system.

2.4 Deep learning

The deep learning era began in 2012. It is a development of machine learning; however, to make a correct prediction, deep learning requires a large number of data, while machine learning needs less data. Deep learning is machine learning algorithms with more than two layers of neural networks that are utilized for complexity modeling of relationships and concepts [5, 34]. Recently, deep learning has been utilized in multiple areas and achieved a high level of performance in areas such as image recognition, filtering social networks, and speech recognition for performing various tasks, including classification, clustering, detection, dimension reduction, and pattern recognition [11].

In the field of computer vision, deep learning is extremely useful. Different applications of deep learning have been utilized with computer vision, including image classification, text recognition, object detection, gender classification, and facial recognition [20]. Based on the architecture, deep learning techniques are categorized into the three main categories [35], which are outlined below.

Generative deep learning

Generative deep learning, or unsupervised deep learning architectures, is utilized with unlabeled data. Recurrent Neural Networks (RNN) is an example of generative architecture.

Discriminative deep learning

Discriminative deep learning, or supervised deep learning architectures, is mainly utilized for pattern classification. CNN is the most common method in this category. CNN comprises several layers that perform feature extraction then classification. The layers of CNNs are composed of input, output, and one or more hidden layers. The hidden layers are convolutional, pooling, normalization, and optional connected layers (see Fig. 5) [27]. Many weights are shared by the convolutional layer. The dimensionality of the features that are extracted by the convolutional layer is reduced by the pooling layer [8]. Compared with other standard feed-forward neural networks that have layers of the same size, CNNs are easy to train and learn as they have fewer parameters and connections. Many studies [15] have shown that CNNs are the commonly used state-of-the-art architectures in the field of computer vision as they are highly effective. In the LSVRC-2010 competition, the researcher in [15] utilized CNN to classify 1.2 million images in the ImageNet

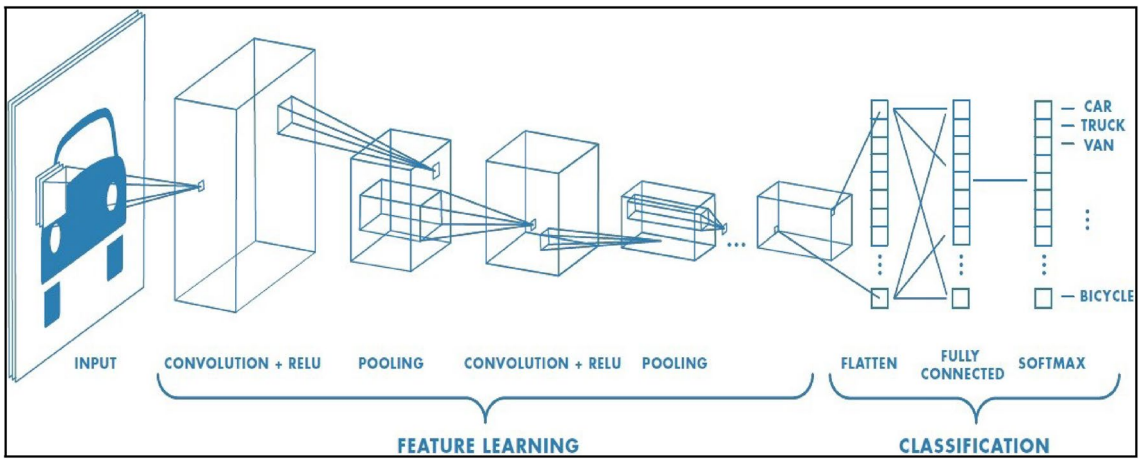


Fig. 5 CNN (neural network with many convolutional layers). (*Source*: [27])

dataset into around 1000 different classes with a minimum error rate.

Hybrid deep learning

Hybrid deep learning, or semi-supervised architecture, is the integration of generative and discriminative architectures. It takes the advantage in the early phases of the generative architecture and in the later stages of the discriminative architecture to recognize data. The most popular method in this category is Deep Neural Network (DNN).

2.4.1 CNN in Image Classification

In computer vision tasks, CNN is particularly useful in image classification [35]. The first time CNN was utilized was in the research of [29] for handwritten character recognition. ImageNet is a large-scale image dataset that can be used in the image classification and object detection research field [9]. It utilizes the WordNet hierarchy structure, with every node in this hierarchy represented by hundreds or thousands of images [37].

The top CNN image classification models that were trained on ImageNet are AlexNet, VGGNet, and Inception. AlexNet was conducted by the researchers in [27]. They trained their CNN on a subset of ImageNet for 6 days and achieved good results compared to other studies conducted at that time. The final error rate was 15.3% with about 60 million parameters. VGGNet was developed by the researchers in [38]. Similar to AlexNet, VGGNet was also trained on ImageNet. The number of parameters was 138 million. VGGNet achieved a 7.3% error rate. The Inception model is a very deep network compared with AlexNet and VGGNet. It was developed by Google. The Inception model achieved a 6.67% error rate. It employed only five million parameters, which is much smaller than the number of parameters utilized in AlexNet and VGGNet. This allows for high-quality training [40].

2.4.2 CNN in Text Classification

In recent years, deep learning models have achieved noteworthy results in computer vision [27]. The researchers in [22] studied the CNN situation to work with text classification, since CNN is commonly used in computer vision tasks. They utilized their findings to suggest improvements of CNNs to be used for text. The researcher in [26] presented some experiments with CNN built on top of pre-trained word2vec for sentence-level text classification tasks. Word2vec vectors were trained on more than 100 million words in a dataset of Google News by using the continuous bag-of-words architecture [31]. The researchers in [3] trained the CNN model on two different datasets on top of word embedding. They employed CNN in classifying an open domain question answering system. The results were promising. Applying CNNs to text classification has shown that convolutional networks without any knowledge of the syntactic or semantic structures of the languages can be applied to different sets of words [47].

Although recent studies on social media monitoring and analysis achieved good results, they failed to incorporate threat detection based on images and text content together. All the research studies discussed above focused on the social media monitoring and tracking; however, in this research, the objective is different. This research attempts to predict whether image content posted on social media contains any threat. Moreover, instead of only text information, this research will utilize the image content in addition to the text content.

To the best of the researchers' knowledge, there is no similar research based on image content in threat detection

on social media. This research explores the existing image classification and object detection techniques related to deep learning. Therefore, it can fill a significant gap in this area of research.

3 Methodology

The experiment for threat detection is conducted utilizing the CNN model on the TensorFlow platform. TensorFlow is a machine learning open-source platform developed by Google to support deep learning algorithms. TensorFlow assists in robust training and can run trained models in multiple platforms. Because of TensorFlow's flexibility, research and experimentation with new models are supported [1]. To achieve the aims of the research, the methodology contains three main phases: data collection, data preprocessing, and classification (see Fig. 6).

3.1 Data collection

This research focuses on pure images without any text embedded in images and posted comments in Arabic separately. In the data collection stage, posts were collected from Instagram utilizing different tools, such as the Instagram API, which was available and open before April 2018. The new API platform was released in December 2018, and therefore, subsequently, the additional data were collected from Google search and changed to the same format. They were extracted to a Comma Separated Values (CSV) file with UTF-8 Unicode to support Arabic text. From the Instagram posts, one dataset was collected manually that consists of 1000 images and more than 2000 Arabic comments from different accounts on different subjects.

3.2 Preprocessing

In the past, researchers utilized different methods for Arabic text preprocessing [4, 13, 18, 21, 33] and Arabic text classification [2, 6, 36]. In this work, the author uses manual preprocessing, which is explained in the steps outlined below.

Cleaning for comments

- Deleted the non-Arabic comments for the sake of simplicity
- Ensured that each comment contains enough words for the purpose of classification
- Deleted all @mention
- Removed punctuation
- Imported tokenizer from keras.preprocessing.text.

Cleaning for images

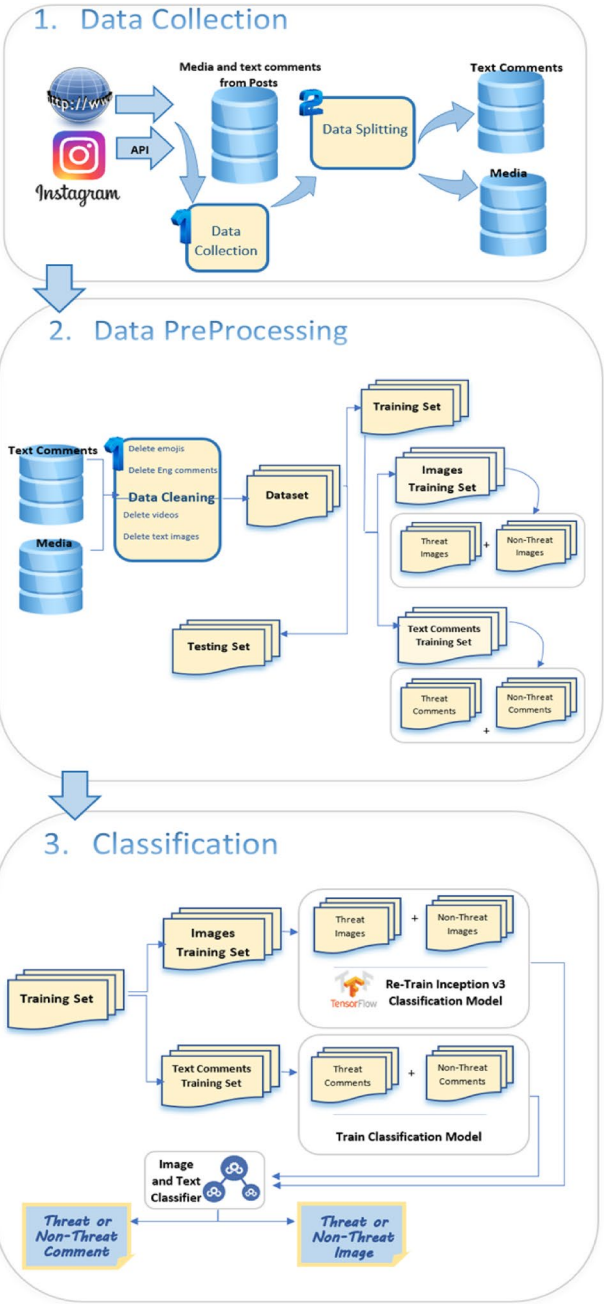


Fig. 6 The methodology of the research

- Deleted selfie images
- Deleted any images that contain only text.

3.2.1 Data labeling

Manual data labeling was adopted instead of automatic data labeling algorithms to achieve more accurate classification. The dataset comments and images were labeled in two classes (threat or non-threat) to create the training set. There were around 2000 comments: more than 1000 were

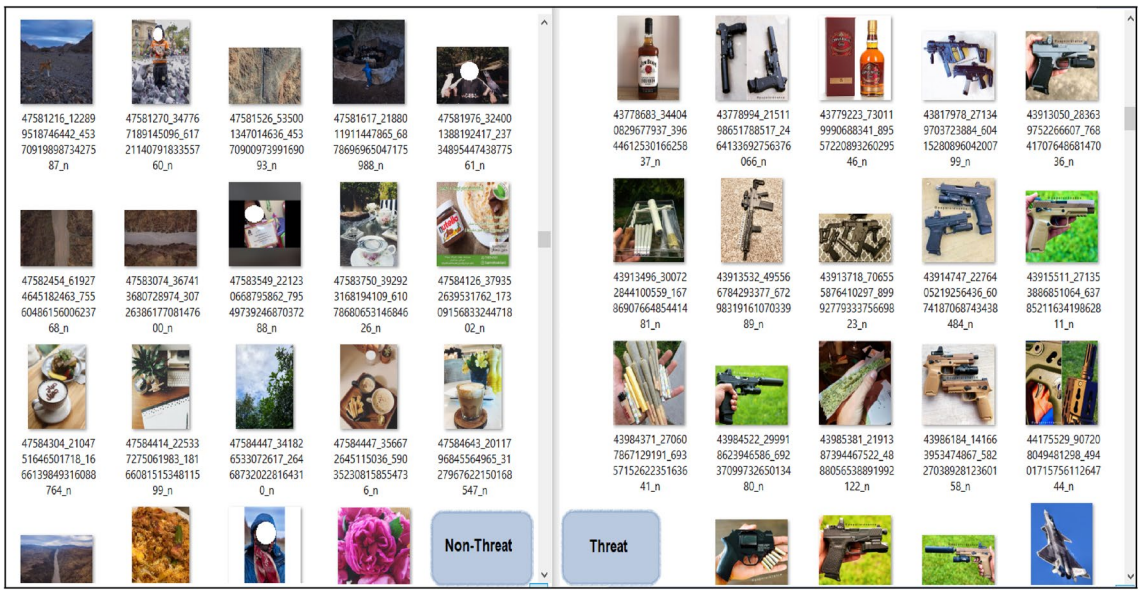


Fig. 7 Examples of threat and non-threat dataset images

threat comments, and the rest were non-threat comments. In addition, there were 1000 images: more than 600 were threat images, and the rest were non-threat images. Threat and non-threat dataset image examples are shown in Fig. 7.

3.3 Classification

The purpose of this research is to create a model that can be utilized in image and text classification. This research is based on the TensorFlow framework. TensorFlow is an open-source platform developed to test with new models and then train them on different datasets [1].

3.3.1 Image classification

In the image classification, the Inception v3 model was utilized in this research. It is a deep CNN that was pre-trained on the ImageNet dataset, which contains more than 1000 different classes [40]. To make the training task easier and faster, transfer learning can be utilized to take advantage of the existing pre-trained models [45]. Since doing the training from scratch requires a large amount of computation power and a huge labeled dataset, the researcher used the transfer learning technique by utilizing the Inception v3 model and performing the training on the collected dataset related to threat [46]. The dimension of the convolution filter was n_dim = 100 and the experiments were carried by taking the learning rate 0.01. The number of training, validation and testing samples utilized for the experimental purpose of image classification was 700, 150, and 150, respectively. Figures 8 and 9 show the accuracy and cross-entropy loss during transfer learning based on the collected image dataset. Validation accuracy is the percentage of the correctly detected samples from a random selection that was not in the original training dataset. Cross-entropy is a loss function that provides an insight into how the learning

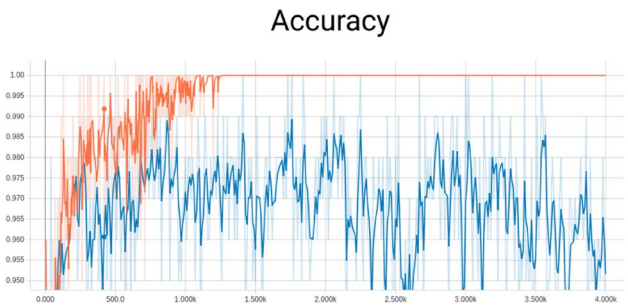


Fig. 8 The accuracy on the Threat, Non-threat image dataset

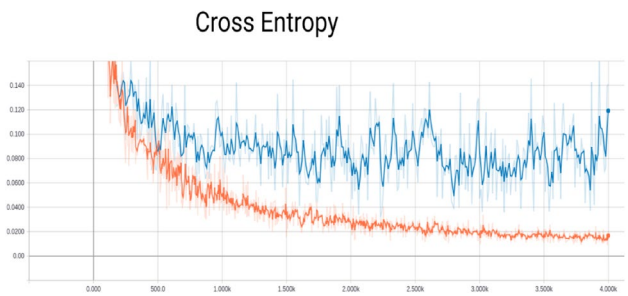


Fig. 9 The cross-entropy on the threat, non-threat image dataset

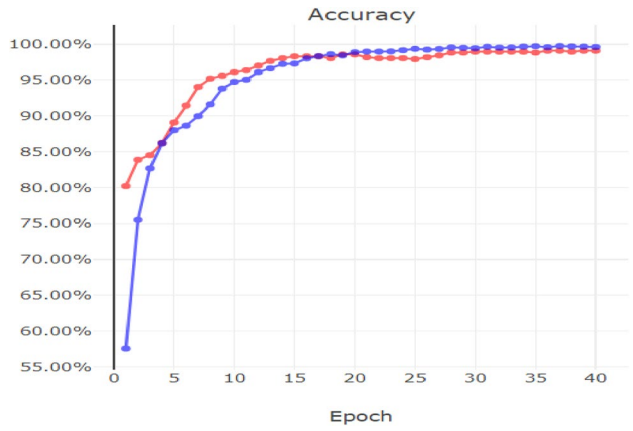


Fig. 10 The accuracy on the threat, non-threat comment dataset

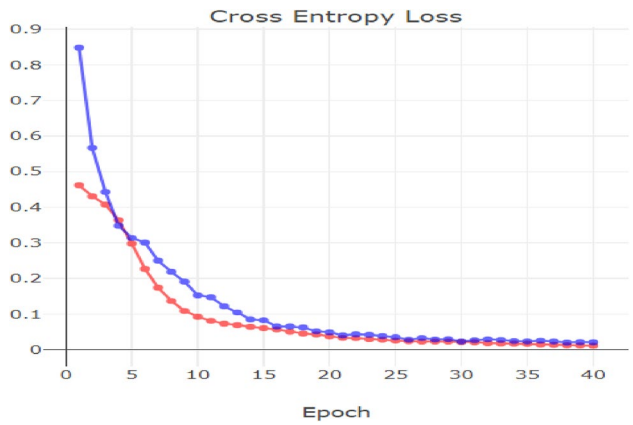


Fig. 11 The cross-entropy on the threat, non-threat comment dataset

is progressing; lower numbers are better [16, 17]. The orange line represents the training set, and the blue line represents the validation set.

3.3.2 Comment classification

The criteria for text classification in CNN are the same as those for image classification; the only difference is that instead of pixel values, the matrix of word vectors is used. The model was developed utilizing the Keras framework that employs TensorFlow as the backend. First, parameters such as how many unique words and their length were set up. Then, the tokenizer was imported from Keras, and tokenization and padding were carried out. Subsequently, the datasets were split into train, validation and test datasets utilizing the function train_test_split imported from sklearn.model_selection. The number of training, validation and testing samples utilized for the experimental purpose of comment classification was 1792, 384, and 384, respectively. Keras embedding layer was used for word embedding on text data (word2vec) [26]. Figures 10 and 11 show the accuracy and cross-entropy loss in the training and validation of the comment dataset, respectively. The training data set is illustrated in blue, and the validation dataset is depicted in red.

4 Experiment results

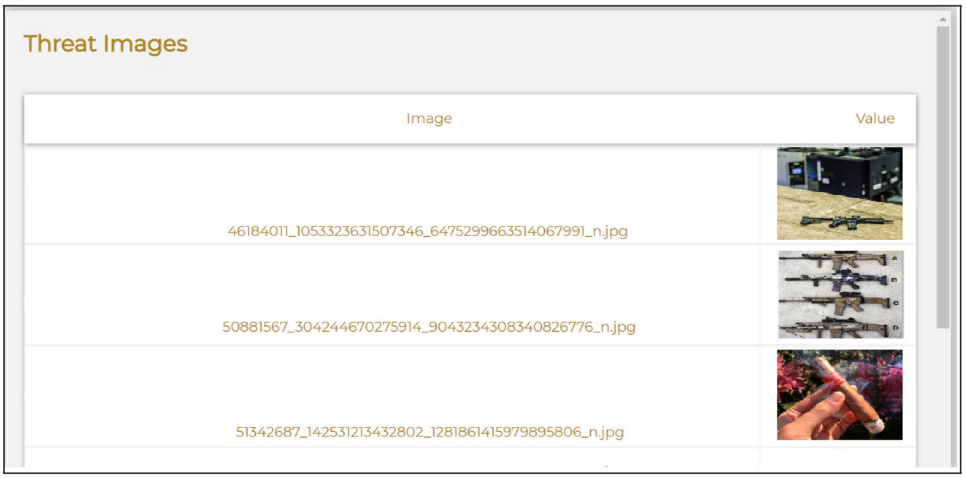
This research was based on the TensorFlow framework and the hardware platform was a Huawei laptop: processor 2.90 GHz Intel i7, RAM 8 GB.

Figure 12 shows the three images detected as threat images, and Fig. 13 shows non-threat images.

Figure 14 shows the threat comments marked as 1 in column E in the below figure, which means they are threat comments. Figure 15 is the same as Fig. 14, but the comments are translated in column F for better understandability of non-Arabic speakers.

To verify whether the models achieve their objectives and aims, the results were evaluated. The confusion matrix was utilized to calculate the number of predicted images

Fig. 12 Screenshot of detected threat images



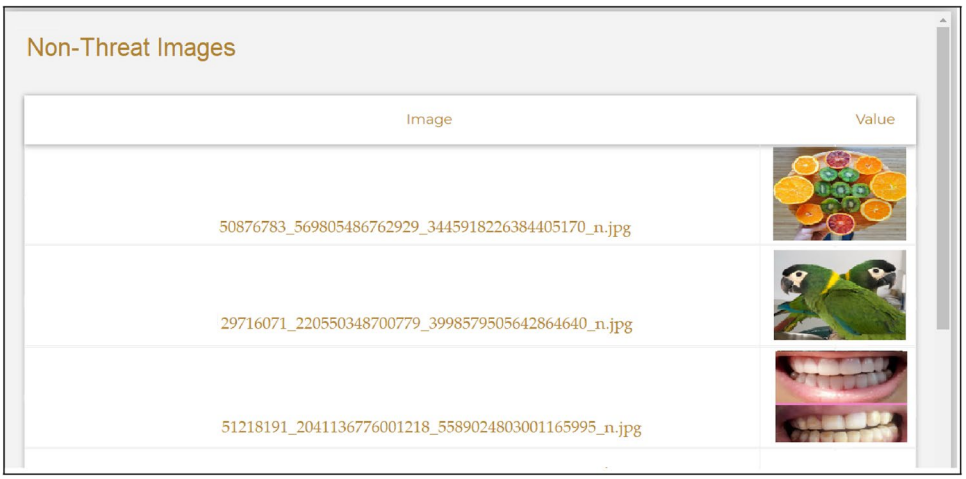
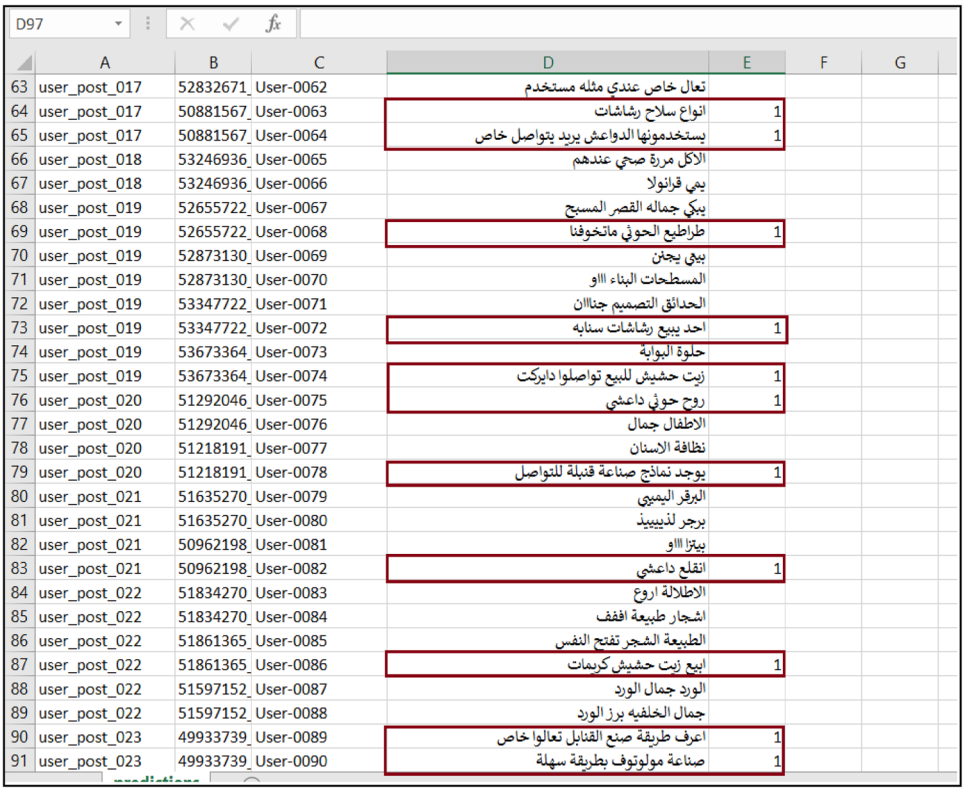


Fig. 13 Screenshot of non-threat images



	A	B	C	D	E
63	user_post_017	52832671_	User-0062	تعال خاص عندي مثله مستخدم	
64	user_post_017	50881567_	User-0063	انواع سلاح رشاشات	1
65	user_post_017	50881567_	User-0064	يستخدمونها الدواعش يريد يتواصل خاص	1
66	user_post_018	53246936_	User-0065	الاكل مررة صحي عندهم	
67	user_post_018	53246936_	User-0066	بعي قرانولا	
68	user_post_019	52655722_	User-0067	يبكي جماله القصر المسبح	
69	user_post_019	52655722_	User-0068	طراطيع الحوثي ماتخوفنا	1
70	user_post_019	52873130_	User-0069	بيبي يجنن	
71	user_post_019	52873130_	User-0070	المسطحات البناء اااو	
72	user_post_019	53347722_	User-0071	الحدائق التصميم جناان	
73	user_post_019	53347722_	User-0072	احد يبيع رشاشات سنابه	1
74	user_post_019	53673364_	User-0073	حلوة البوابة	
75	user_post_019	53673364_	User-0074	زيت حشيش للبيع تواصلوا دايركت	1
76	user_post_020	51292046_	User-0075	روح حوثي داعشي	1
77	user_post_020	51292046_	User-0076	الاطفال جمال	
78	user_post_020	51218191_	User-0077	نظافة الاسنان	
79	user_post_020	51218191_	User-0078	يوجد نماذج صناعة قنبلة للتواصل	1
80	user_post_021	51635270_	User-0079	البرقر البمبي	
81	user_post_021	51635270_	User-0080	برجر لذيييذ	
82	user_post_021	50962198_	User-0081	بيتزا اااو	
83	user_post_021	50962198_	User-0082	انقلع داعشي	1
84	user_post_022	51834270_	User-0083	الاطلالة اروع	
85	user_post_022	51834270_	User-0084	اشجار طبيعة اففف	
86	user_post_022	51861365_	User-0085	الطبيعة الشجر تفتح النفس	
87	user_post_022	51861365_	User-0086	ابيع زيت حشيش كريمات	1
88	user_post_022	51597152_	User-0087	الورد جمال الورد	
89	user_post_022	51597152_	User-0088	جمال الخلفيه برز الورد	
90	user_post_023	49933739_	User-0089	اعرف طريقة صنع القنابل تعالوا خاص	1
91	user_post_023	49933739_	User-0090	صناعة مولوتوف بطريقة سهلة	1

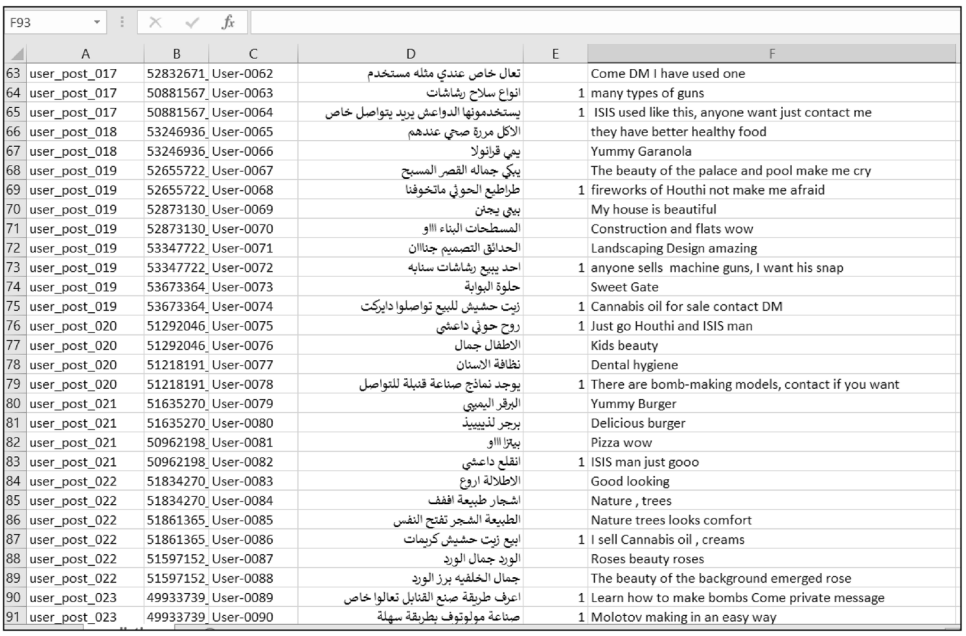
Fig. 14 Screenshot of detected threat comments

and comments that were counted as true positive, true negative, false negative, and false positive. The confusion matrix [10] can be described as a table that shows the classification model performance of a set of test data of which the true values are already known, as shown in Table 1. The basic terms in the confusion matrix are as follows:

- *True positives (TP)* The number of instances predicted to be positive, and they are really positive, so they are classified correctly.
- *True negatives (TN)* The number of cases predicted to be negative, and they are really negative, so they are classified correctly.
- *False positives (FP)* The number of instances predicted to be positive, but they are actually negative instances.
- *False negatives (FN)* The number of cases predicted to be negative, but they are actually positive cases.

The following is a list of parameters that are computed and derived from the confusion matrix for any classifier:

Fig. 15 Translated comments



	A	B	C	D	E	F
63	user_post_017	52832671_	User-0062	تعال خاص عندي مثله مستخدم		Come DM I have used one
64	user_post_017	50881567_	User-0063	انواع سلاح رشاشات	1	many types of guns
65	user_post_017	50881567_	User-0064	يستخدمونها الدواعش يريد يتواصل خاص	1	ISIS used like this, anyone want just contact me
66	user_post_018	53246936_	User-0065	الاكل مررة صحي عندهم		they have better healthy food
67	user_post_018	53246936_	User-0066	يمي قرانولا		Yummy Garanola
68	user_post_019	52655722_	User-0067	يبكي جماله القصر المسبح		The beauty of the palace and pool make me cry
69	user_post_019	52655722_	User-0068	طراطيع الحوثي ماتخوفنا	1	fireworks of Houthi not make me afraid
70	user_post_019	52873130_	User-0069	بيتي يحلن		My house is beautiful
71	user_post_019	52873130_	User-0070	المسطحات البناء ااو		Construction and flats wow
72	user_post_019	53347722_	User-0071	الحدائق التصميم جناان		Landscaping Design amazing
73	user_post_019	53347722_	User-0072	احد يبيع رشاشات سنابه	1	anyone sells machine guns, I want his snap
74	user_post_019	53673364_	User-0073	حلوة البوابة		Sweet Gate
75	user_post_019	53673364_	User-0074	زيت حشيش للبيع تواصلوا دايركت	1	Cannabis oil for sale contact DM
76	user_post_020	51292046_	User-0075	روح حوثي داعشي	1	Just go Houthi and ISIS man
77	user_post_020	51292046_	User-0076	الاطفال جمال		Kids beauty
78	user_post_020	51218191_	User-0077	نظافة الاسنان		Dental hygiene
79	user_post_020	51218191_	User-0078	يوجد نماذج صناعة قنبلة للتواصل	1	There are bomb-making models, contact if you want
80	user_post_021	51635270_	User-0079	البرقر البمبي		Yummy Burger
81	user_post_021	51635270_	User-0080	برجر لذيييذ		Delicious burger
82	user_post_021	50962198_	User-0081	بيتزا اااو		Pizza wow
83	user_post_021	50962198_	User-0082	انقلع داعشي	1	ISIS man just gooo
84	user_post_022	51834270_	User-0083	الاطلالة اروع		Good looking
85	user_post_022	51834270_	User-0084	اشجار طبيعة افففف		Nature , trees
86	user_post_022	51861365_	User-0085	الطبيعة الشجر تفتح النفس		Nature trees looks comfort
87	user_post_022	51861365_	User-0086	ابيع زيت حشيش كريمات	1	I sell Cannabis oil , creams
88	user_post_022	51597152_	User-0087	الورد جمال الورد		Roses beauty roses
89	user_post_022	51597152_	User-0088	جمال الخلفيه برز الورد		The beauty of the background emerged rose
90	user_post_023	49933739_	User-0089	اعرف طريقة صنع القنابل تعالوا خاص	1	Learn how to make bombs Come private message
91	user_post_023	49933739_	User-0090	صناعة مولوتوف بطريقة سهلة	1	Molotov making in an easy way

Table 1 Confusion matrix

	Predicted Class	
	Negative	Positive
Actual class		
Negative	TN	FP
Positive	FN	TP

- Accuracy, which is the overall success rate:

$$Accuracy = \frac{TP + TN}{Total\ Instances} \times 100 \quad (1)$$

- Misclassification rate (error rate), which means how often the classifier is wrong:

$$Error\ rate = \frac{FP + FN}{Total\ Instances} \quad (2)$$

- True positive rate (TPR), which is the percentage of positive instances predicted correctly, also known as "sensitivity" or "recall:"

$$Recall = \frac{TP}{Actual\ Positive\ Instances} \quad (3)$$

- Precision, which is the percentage of instances marked as positive and they are actually positive:

$$Precision = \frac{TP}{Prdicted\ Positive\ Instances} \quad (4)$$

Table 2 Calculations of measures for image and comment detection

Measure	Calculated value	
	Image	Comment
Accuracy	0.96	0.99
Misclassification rate (error rate)	0.04	0.01
True positive rate (recall)	0.92	0.99
False positive rate	0.03	0.01
True negative rate (specificity)	0.97	0.98
Precision	0.92	0.99
Prevalence	0.26	0.55
F-score	0.92	0.99

- F-score, which is a weighted average of the precision and TPR (recall):

$$F - Score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \quad (5)$$

This research is related to Instagram posts and mainly focuses on threat detection in images and comments posted by users. Each post contains at least one image with comments. Most of the time, the users discuss the image they posted in their comment (related), and sometimes the comments are not related to the image. Therefore, the classification is for the image and comments for that post, as shown separately in Table 2. The calculations in Table 2 derived from the confusion matrix demonstrated that the proposed technique for the image classification into threat and non-threat images that was retrained utilizing the Inception v3 model achieved a 96% accuracy rate, whereas the proposed technique for the comment classification into threat and

non-threat comments, which is the CNN model, achieved an overall accuracy rate of 99%.

5 Conclusion

A significant amount of attention has been paid to deep learning for text classification along with computer vision tasks. Those tasks are enhanced by investigating the need for techniques for social media content analysis and tracking. Instagram is a popular social media website and mobile application, where users are allowed to upload, view, share, and comment on posts. It offers a wide variety of post content.

Analysis and tracking of social media content have been investigated by many studies in the past. However, to the best of the researchers' knowledge, no research has focused on threat detection in Instagram image posts, especially for Arabic comments. To fill this research gap, this research aims to detect threats in Instagram image posts and Arabic comments by utilizing some deep learning algorithms. In this research, an image and text classifier model is built to detect threats in Instagram posts. The results show that the developed image classifier achieved 96% accuracy, and the comments classifier achieved 99% accuracy. This research work can be extended by classifying the posts based on the context of the conversation and the occurrence of any threat.

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X (2016) TensorFlow: a system for large-scale machine learning, pp 265–283
2. Adel HM (2019) Arabic text classification: a review. Modern Appl Sci. https://doi.org/10.5539/mas.v13n5p88
3. Amin MZ, Nadeem N (2018) Convolutional neural network: text classification model for open domain question answering system. arXiv preprint arXiv:1809.02479
4. Ayah S (2019) Deep learning for sentiment analysis of arabic text. In: ArabWIC 2019: proceedings of the ArabWIC 6th annual international conference, pp 1–8
5. Bengio Y, Goodfellow IJ, Courville A (2015) Deep learning. Nature 521:436–444
6. Biniz M, Boukil S, Adnani F, Cherrat L, Moutaouakkil A (2018) Arabic Text classification using deep learning technics. Int J Grid Distrib Comput 11:103–114
7. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, San Diego, CA, USA, pp 886–893
8. Deng L, Yu D (2014) Deep Learning: Methods and Applications. Found Trends® Signal Process 7:197–387. https://doi.org/10.1561/2000000039
9. Deng J, Dong W, Socher R, Li L-J, Li Kai, Fei-Fei Li (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, Miami, FL, pp 248–255
10. Dietrich D, Heller B, Yang B, EMC Education Services (2015) Data science & big data analytics: discovering, analyzing, visualizing and presenting data. Wiley, Indianapolis, IN
11. Dong B, Wang X (2016) Comparison deep learning method to traditional methods using for network intrusion detection. In: 2016 8th IEEE international conference on communication software and networks (ICCSN). IEEE, Beijing, China, pp 581–585
12. Geethu GS, Kamatchi T (2016) Recognition of facial expressions in image sequence using multi-class SVM. Int J Innov Res Comput Commun Eng 4(8):14630–14638
13. Ghadah A, Taha O, Thomas HR (2017) Challenges in sentiment analysis for arabic social networks. Proc Comput Sci 117:89–100
14. Hadid A, Member S (2006) Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Anal Mach Intell 28(12):2037–2041
15. Hassannejad H, Matrella G, Ciampolini P, De Munari I, Mordonini M, Cagnoni S (2016) Food Image recognition using very deep convolutional networks. In: Proceedings of the 2nd international workshop on multimedia assisted dietary management. ACM, New York, NY, USA, pp 41–49
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, Las Vegas, NV, USA, pp 770–778
17. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv :1704.04861 Cs
18. Imane G, Houda S, Faical A, Billel G, Damien N (2019) Arabic natural language processing: an overview. J King Saud Univ Comput Inf Sci (in Press, available online 23 February 2019)
19. Instagram A (2019) Instagram: active users 2018| Statista. https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/. Accessed 18 Nov 2019
20. Islam SMS, Rahman S, Rahman MM, Dey EK, Shoyaib M (2016) Application of deep learning to computer vision: A comprehensive study. In: 2016 5th international conference on informatics, electronics and vision (ICIEV). pp 592–597
21. Itani M, Roast CR, Al-Khayatt S (2017) Corpora for sentiment analysis of Arabic text in social media, pp 64–69
22. Jacovi A, Sar Shalom O, Goldberg Y (2018) Understanding convolutional neural networks for text classification. In: Proceedings of the 2018 EMNLP workshop blackboxNLP: analyzing and interpreting neural networks for NLP. Association for Computational Linguistics, Brussels, Belgium, pp 56–65
23. Johansson F, Brynielsson J, Quijano MN (2012) Estimating citizen alertness in crises using social media monitoring and analysis. In: 2012 European intelligence and security informatics conference. IEEE, Odense, Denmark, pp 189–196
24. Kandias M, Stavrou V, Bozovic N, Gritzalis D (2013) Proactive Insider threat detection through social media: the youtube case. In: Proceedings of the 12th ACM workshop on workshop on privacy in the electronic society. ACM, New York, NY, USA, pp 261–266
25. Kang X, Li S, Benediktsson JA (2014) Feature extraction of hyperspectral images with image fusion and recursive filtering. IEEE Trans Geosci Remote Sens 52:3742–3752. https://doi.org/10.1109/TGRS.2013.2275613
26. Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1746–1751
27. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural

information processing systems, v ol 25. Curran Associates, Inc., pp 1097–1105
28. Kumar G, Bhatia PK (2014) A detailed review of feature extraction in image processing systems. In: 2014 fourth international conference on advanced computing & communication technologies. IEEE, Rohtak, India, pp 5–12
29. LeCun Y, Bengio Y (1998) Convolutional networks for images, speech, and time-series. The handbook of brain theory and neural networks, pp 255–258
30. Michel P, Kaliouby RE (2003) Real time facial expression recognition in video using support vector machines. In: ICMI'03: Proceedings of the 5th international conference on Multimodal interfaces, pp 258–264
31. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, vol 26. Curran Associates, Inc., pp 3111–3119
32. Nadeem MSM (2018) Text mining and social media analysis of pizza industry using R. 4:7
33. Noor AM, Sandra K (2018) Preprocessing does matter: parsing non-segmented Arabic. In: Proceedings of the 17th international workshop on Treebanks and linguistic theories (TLT 2018), December 13–14, 2018, Oslo University, Norway
34. Patterson J, Gibson A (2017) Deep learning: a practitioner's approach. O'Reilly Media Inc, Sebaastopol
35. Ponti MA, Ribeiro LSF, Nazare TS, Bui T, Collomosse J (2017) Everything you wanted to know about deep learning for computer vision but were afraid to ask. In: 2017 30th SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T). IEEE, Niterói, pp 17–41
36. Qadi LA, Rifai HE, Obaid S, Elnagar A (2019) Arabic text classification of news articles using classical supervised classifiers. In: 2nd international conference on new trends in computing sciences (ICTCS), Amman, Jordan, 2019, pp 1–6
37. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. arXiv :1409.0575 Cs
38. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 Cs
39. Sykora MD, Jackson TW, OBrien A, Elayan S (2013) National security and social media monitoring: a presentation of the EMOTIVE and related systems. In: 2013 European intelligence and security informatics conference. IEEE, Uppsala, Sweden, pp 172–175
40. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. arXiv :1512.00567 Cs
41. Thai LH, Hai TS, Thuy NT (2012) Image classification using support vector machine and artificial neural network. Int J Inf Technol Comput Sci 4:32–38. https://doi.org/10.5815/ijitcs.2012.05.05
42. Tian D (2013) A review on image feature extraction and representation techniques. Int J Multimed Ubiquitous Eng 8(4):385–395
43. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition (CVPR 2001). IEEE Comput. Soc, Kauai, HI, USA, pp I-511-I–518
44. Whitehill J, Omlin CW (2006) Haar features for FACS AU recognition. In: 7th international conference on automatic face and gesture recognition (FGR06). pp 5–101
45. Wu Y, Qin X, Pan Y, Yuan C (2018) Convolution neural network based transfer learning for classification of flowers. In: 2018 IEEE 3rd international conference on signal and image processing (ICSIP), pp 562–566
46. Xia X, Xu C, Nan B (2017) Inception-v3 for flower classification. In: 2017 2nd international conference on image, vision and computing (ICIVC). IEEE, pp 783–787
47. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates, Inc., pp 649–657