Journal of Engineering and Technology

ISSN 2180-3811 eISSN 2289-814X https://jet.utem.edu.my/jet/index

DETECTING CORONAVIRUS DISEASE OF ARABIC TWEETS USING SENTIMENT ANALYSIS

A. R. Alobaidi*[1] and Z. N. Nuimi[2]

[1] Department of Computer Techniques Engineering, Faculty of Information Technology, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq.

[2] Department of Electrical Engineering, Faculty of Engineering, University of Mayasan, Mayasan, Iraq.

corresponding_abdulazeez.riyadh@sadiq.edu.iq

Article history:
Received Date: 18 September 2022
Revised Date: 29 November 2022
Accepted Date: 15 December 2022

Keywords:
Machine Learning, COVID-19, Arabic Tweets, Twitter, Hold-out.

Abstract— In order to detect COVID-19 from Arabic tweets, we used five machine learning algorithms which are Support Vector Machine, Decision Tree, Multinomial Naive Bayes, Random Forest, and Voting Classifier. We have collected and analysed COVID-19 tweets written in the Arabic language. We used the Hold Out method with different test sizes to split the dataset into training and testing datasets. The obtained results show that the Voting Classifier is the best classifier with a 94.25% accuracy in the test size of 0.5 for the detection of COVID-19 from Arabic tweets.

I. Introduction

The spread of using social media, such as Twitter and Facebook, has significantly increased due to the phenomenal development of technology and the ease of accessing the internet on any device [1]. Users of social media, particularly Twitter, can post or publish anything that happens in their daily lives, such as political issues, economics, or any sick symptoms like infectious diseases felt by the patient [2].

Many diseases and disorders afflict a person with many symptoms, which may lead to death in many cases [3]. Infectious diseases are disorders caused by pathogenic microorganisms like bacteria, viruses, and parasites [3]. There are several ways to spread these diseases among people, directly or indirectly, from person to person, animals, or by consuming contaminated food or water [3].

COVID-19 is a serious infectious disease that appeared in China in late 2019 and then spread to most countries in the world. It is considered a serious disease due to its rapid spread among humans [4]. In March 2020, this virus was classified as a global epidemic by the World Health Organization because it has infected millions of people [4].

In the proposed work, we collected Arabic tweets using Arabic keywords to detect COVID-19 epidemics. Several machine learning algorithms are applied to the data to discover if COVID-19 can be detected through the study of Arabic tweets. These tweets are classified manually as positive or negative by human experts. The main goals of this paper are: 1) to study the reactions of Arab people to this virus, whether they are serious about it or if it is a joke. 2) to determine whether the disease has spread throughout the Arab world. 3) to find out if people are satisfied with the measures taken by their country to confront the virus. 4) be aware of the available prevention methods to protect oneself. 5) recognize the most common causes of death.

The rest of the paper is organized as follows: Section 2 describes the related works. Section 3 presents the workflow of the proposed system. The experiments and the obtained results are presented and discussed in Section 4 with some future recommendations. Section 5 concludes this paper with a summary and an outlook towards future works.

II. Related Works

There are a lot of studies that apply sentiment analysis using several algorithms to detect infectious diseases such as Ebola, SARS, etc; using social media data languages. Among these studies, we can cite the system of Baker et al. [5] which studied epidemic diseases in the Arab world as the first study to collect Arabic tweets. These tweets are related to the influenza outbreaks in the Arab world.

Joshi et al. [6] presented two experiments related to the Ebola epidemic that spread in West Africa. Samaras et. al. [7] built a model with two systems to monitor influenza through the test period. These systems are the search engine and social network data. Barman et al. [8] evolved machine learning algorithms to determine the infectious disease and the host genes that are related to it. This is done by merging the sequence and protein features.

Chae et al. [9] collected datasets from search query data and social media data. Then, they applied three deep learning models to detect infectious diseases in general: deep neural network, autoregressive integrated moving average, and Long-Short Term Memory.

III. Methodology

This paper focuses on proposing a new system to detect the COVID-19 disease from Arabic Tweets. Firstly, we classify the data into positive and negative. Positive refers to tweets related to this disease, and negative refers to tweets not related to this disease. Secondly, we applied the preprocessing steps to the classified dataset to make it clean, error-free, and understandable using many tools like Java, Excel, and

Python. Thirdly, we used machine learning algorithms which are Support Vector Machine, Decision Tree, Multinomial Naive Bayes, Random Forest, and Voting Classifier. Finally, we compared the used machine learning algorithms based on many metrics, such as precision, accuracy, f1-score, and recall. The proposed system in this paper falls into the detection of coronavirus disease in Arabic tweets using sentiment analysis as shown in Figure 1.

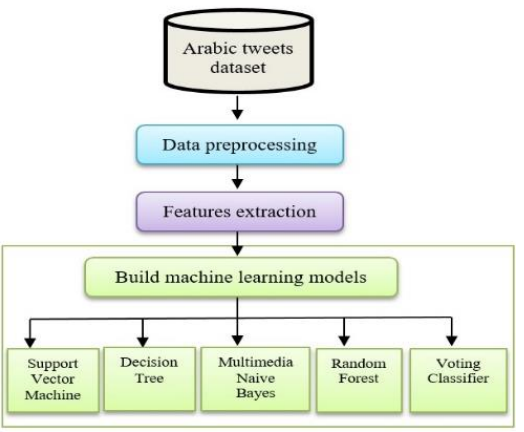

Figure 1: General Architecture of the proposed system of detecting coronavirus disease in Arabic tweets using sentiment analysis

A. Dataset

We collected 24,690 Arabic tweets out of 84,956 tweets using 33 Arabic keywords like, 19- كوفيد" COVID-19", "وباء الكورونا, Corona Epidemic", and " ، القاتل الفيروسKiller Virus". These keywords are obtained from Covid-19 symptoms and medical sites. The reason behind the number of tweets decreasing is that there are a lot of duplicate tweets.

The process of collecting this data involves many steps. The first step should be to create a Twitter account [10]. The second step is to create an account on Twitter Application Management [11] to get the Twitter API keys. The third step is to download the Twitter library, which is called

Twittre4J [12]. So, the keys and the library are used in Java code to allow us to collect this data from Twitter. In this code, we determine the keyword in the Arabic language, and we define the number of retrieved tweets in each run. After that, the data will be stored in the text file. Then, we applied on text file the pre-processing steps to make the data ready to be put into an Excel file. Finally, we classify the data into two categories that are positive and negative. More details are presented in the next subsection.

B. Categories

The data is manually annotated by human experts into two categories which are positive and negative, as shown in Figure 2.

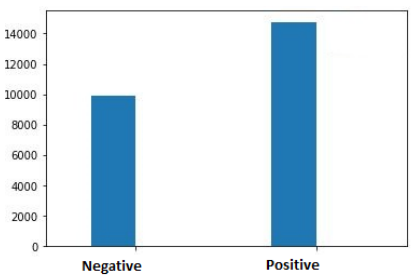

Figure 2: Number of positive and negative examples

The validation of annotation is done by other human experts not participated in the first task of annotation.

- Positive Tweet:

The positive tweets indicate that the tweets are related to the virus and that the user is suffering from symptoms. The number of positive tweets is 14,765 tweets.

- Negative Tweet:

The negative tweets indicate that the tweets are not related to the virus, the user is not suffering from this virus from any symptoms, or the scientist has found a vaccine for the virus. The number of negative tweets is 9,925 tweets.

Through the symptoms that are mentioned in a positive tweet like ("cold" or " or "الزكام" or "المعاناة" or "المرض"), and (" ممكن 19 عندي كوفيد" or " I suffer from coronavirus "), the tweet indicates that the user is suffering from COVID-19. Through the word ("find vaccine" or "أدوية" or " انتاج لقاح") in the negative tweet, the tweet indicates that the user wrote that it was possible to find a vaccine for COVID-19.

C. Data Preprocessing

After collecting the data, it is ready for pre-processing because it is not clean, it contains a lot of things that are not understood, and it is not useful for the classification process. In the proposed work, the pre-processing step is done in three ways. The first way uses Java code to remove tweet duplicates. The second way is by using the Excel tool to remove anything in the dataset except Arabic words. The third way using Python is to remove retweet words (RT), special characters, numbers, Twitter user mentions, and hashtags.

D. Features Extraction

In this step, we extract features to feed them into all machine learning algorithms. Since the algorithms of machine learning cannot deal with the text directly, they must convert it into numbers using many techniques. In our work, we used a CountVectorizer technique which is a method used to convert text to numerical data [13]. The CountVectorizer technique makes it easy for text data to be used directly in machine learning such as text classification.

E. Machine Learning Algorithm

The data is fully preprocessed and classified into positive and negative categories. So, we applied several machine learning algorithms to the data to build the models. These models help us to predict the label of unseen data or testing datasets, whether they are positive or negative. The type of these algorithms is classification because each tweet in the dataset has a label. There are a lot of techniques that split the data into training and testing data. We used in our experiment hold-out training method. More precisely, we used the cross-validation hold-out method, where the machine learning model will first train using a portion of data, and then it will be tested on what's left. We used different test sizes to split the data into training and testing (0.1, 0.2, 0.3, 0.4, and 0.5). For example, 0.1 means that we used 10% of the data for training and the rest

90% for testing. 0.5 means that we used 50% of the data for training and the rest 50% for testing.

The machine learning algorithms applied to the data are Multinomial Naive Bayes (MNB), the Voting Classifier (VC), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). The details of these algorithms and their parameters are shown in the following subsections:

- Multinomial Naïve Bayes algorithm:

The multinomial Naïve Bayes algorithm is a supervised approach that requires a learning dataset before beginning to work [14]. In our experiments, the parameters are alpha with a value of 1.0 and fit prior with a value of True.

- Decision Tree algorithm:

A decision tree is used for both regression and classification works. It has a hierarchical, tree structure consisting of a root node, branches, internal nodes, and leaf nodes [15]. In our experiments, the parameters are that the criterion is the Gini index, the splitter is the best, and the min_samples_split is equal to 2.

- Support vector machine:

The SVM algorithm is to construct the most suitable line or decision boundary that can segregate n-dimensional space into classes [16]. Many types of Kernel functions are defined like Gaussian and linear kernels. In our proposed work, we used the linear kernel.

- Random Forest classifier:

The random forest is a classification algorithm consisting of many decision trees [17]. In our work, the parameters are the estimators which are equal to 100, the criterion is the Gini index, and the min_samples_split is equal to 2.

- Voting Classifier:

A voting classifier is an estimator which trains different models and predicts based on aggregating the results of each estimator [13]. In our work, we implemented the four classifiers that combined them in the Voting Classifier which is a machine learning estimator that trains various base models or estimators and predicts based on

aggregating the findings of each base estimator. The aggregating criteria can be combined decision of voting for each estimator output. The parameters are voting type, hard, and the base classifier, which combines the RF, NB, DT, and SVM.

IV. Experimental Results and Discussion

A. Experimental Results

Table 1 shows the results of the machine learning algorithms with different test sizes, and we compiled these algorithms based on data mining metrics.

Table 1: Machine learning algorithms results with different test sizes

Size	Metric	SVM	DT	MNB	RF	VC
0.1	Accuracy	77.40	90.50	75.20	91.52	94.00
	Precision	77.00	90.00	75.00	91.00	95.00
	Recall	77.00	90.00	75.00	91.00	94.00
	F1-Score	77.00	90.00	75.00	91.00	94.00
0.2	Accuracy	77.35	90.55	75.44	91.56	93.70
	Precision	77.00	90.00	75.00	91.00	94.00
	Recall	77.00	90.00	75.00	91.00	94.00
	F1-Score	77.00	90.00	75.00	91.00	94.00
0.3	Accuracy	77.31	90.59	75.36	91.58	94.03
	Precision	77.00	90.00	75.00	91.00	95.00
	Recall	77.00	90.00	75.00	91.00	94.00
	F1-Score	77.00	90.00	75.00	91.00	94.00
0.4	Accuracy	77.90	90.62	75.89	91.61	94.30
	Precision	78.00	90.00	76.00	91.00	95.00
	Recall	78.00	90.00	76.00	91.00	94.00
	F1-Score	77.00	90.00	75.00	91.00	94.00
0.5	Accuracy	78.02	90.68	75.84	91.65	94.25
	Precision	78.00	90.00	76.00	91.00	95.00
	Recall	78.00	90.00	76.00	91.00	94.00
	F1-Score	78.00	90.00	75.00	91.00	94.00

These metrics are accuracy, precision, recall, and f1-score. The formula for each of them are shown in Equation (1), (2), (3) and (4) as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FO + FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

$$f1-score = \frac{(2 * Precision * Recall)}{(Precision * Recall)} \quad (4)$$

where:
TP = True positives; for correctly predicted event values
FP = False positives; for incorrectly predicted event values
TN = True negatives; for correctly predicted no-event values
FN = False Negatives; for incorrectly predicted no-event values

As can be seen from Table 1, the best algorithm that gave the best performance based on these metrics is the voting classifier in test size equal to 0.5 with 94.25%, 95%, 94%, and 94% respectively. The results mean that the voting classifier can predict unseen tweets, whether positive or negative, with an accuracy of 94.25%. It is also important to notice that we obtained a precision of 95%, a recall of 94% and an f1-score of 94%.

B. Discussion

In this study, we applied five machine learning algorithms with five test sizes. The best one is the Voting Classifier with a 0.5 test size. We obtained an accuracy (94.25%), precision (95%), recall (94%), and f1-score (94%). The RF and DT are almost the same. The same thing goes for the SVM and MNB. The Voting Classifier is combined between four classifiers and sometimes gives better results than others as shown in Figure 3.

The goals of this work are 1) to study the reactions of Arab people to this virus, whether they are serious about it or joking about it. 2) To determine whether the disease has spread throughout the Arab world. And 3) to find out if people are satisfied with the measures taken by their country to confront the virus. 4) Be aware of the prevention methods available to protect oneself from it. 5) To be aware of the most common cause of death.

For the first goal, the Arab people's reactions are varied. 60% of them said that the virus is dangerous and serious, but the

rest said it is not. It can be seen from the analysis of our data that the major of Arab people are serious about COVID-19.

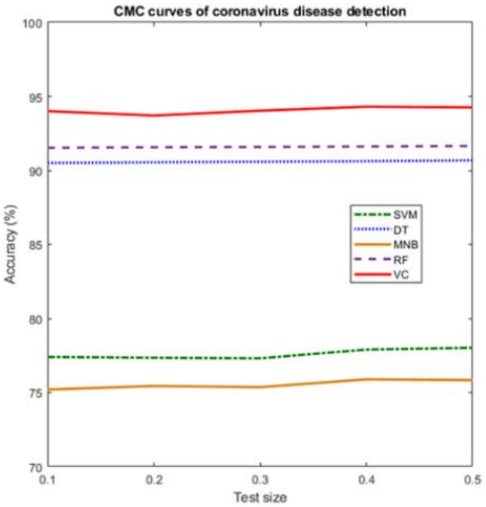

Figure 3: CMC curves of coronavirus disease detection

For the second goal, the virus has spread in all Arab countries, but the number of cases is less than the number of cases in America and Europe.

For the third goal, people are satisfied with the measures that the government has taken. For this reason, the number of cases is down.

For the fourth goal, there are many prevention methods to protect from the virus, such as washing hands very well, not leaving the home, and keeping a safe distance between people and places.

For the last goal, is failure to obey the government concerning procedures and leniency towards them, not washing hands and rapprochement between people in crowded places.

Some recommendations from several governments about COVID–19 to reduce the spread of it are washing hands very well, not rapprochement between people in crowded places, and people must understand the preventive and precautionary measures taken by the government. Do not be reckless about these procedures.

V. Conclusion and Future Works

In this paper, we analyze COVID-19, which happened in China at the end of 2019. The analysis was done by collecting data related to this virus from Twitter using 33 Arabic keywords. Five machine learning algorithms are used with five test sizes (0.1, 0.2, 0.3, 0.4, and 0.5) in Python after applying preprocessing techniques. These algorithms are SVM, DT, MNB, RF, and VC. The best one is the voting classifier with a 0.5 test size with accuracy (94.25%), precision (95%), recall (94%), and f1-score (94%). This result indicates that the machine learning algorithms can detect COVID-19 through Arabic Tweets. The Arab people's reactions are varied. 60% said the virus is dangerous and serious, but the rest said it is not. The virus has spread in all Arab countries, but the number of cases is less than the number of cases in America and Europe. People are satisfied with the measures that the government has taken. For this purpose, the number of cases is down.

In future works, we intend to increase the dataset size and apply Deep Learning (DL) algorithms, not just machine learning, to see if the DL can detect COVID-19 like artificial neural networks, recurrent neural networks, and conventional neural networks. In addition, we intend to use natural language processing tools like Madamira to extract more features (i.e., named entity recognition and part of speech tagging). Finally, we aim to get the reports from the World Health Organization website to compare them with the tweets.

VI. References

[1] Santos, J. C., & Matos, S. (2014). "Analysing Twitter and web queries for flu trend prediction.", *Theoretical Biology and Medical Modelling*, 11(1), 1-11.

[2] Centers for Disease Control and Prevention. (2020). Coronavirus disease 2019 (COVID-19): symptoms of coronavirus. Atlanta: Centers for Disease Control and Prevention.

[3] Infectious Diseases, *EverydayHealth.com*. [Online] Available:

https://www.everydayhealth.com/%20infectious-diseases/

[4] Coronavirus disease 2019 (COVID-19) - Symptoms and causes, *Mayo Clinic*, [Online] Available: https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963

[5] Baker, Q. B., Shatnawi, F., Rawashdeh, S., Al-Smadi, M., & Jararweh, Y. (2020). "Detecting epidemic diseases using sentiment analysis of arabic tweets.", *J. Univers. Comput. Sci*., 26(1), 50-70.

[6] Joshi, A., Sparks, R., Karimi, S., Yan, S. L. J., Chughtai, A. A., Paris, C., & MacIntyre, C. R. (2020). "Automated monitoring of tweets for early detection of the 2014 Ebola epidemic.", *Journal PloS one*, 15(3), e0230322.

[7] Samaras, L., García-Barriocanal, E., & Sicilia, M. A. (2020). "Comparing Social media and Google to detect and predict severe epidemics." *Scientific Reports*, 10(1), 1-11.

[8] Barman, R. K., Mukhopadhyay, A., Maulik, U., & Das, S. (2019). "Identification of infectious disease-associated host genes using machine learning techniques.", *BMC bioinformatics*, 20(1), 1-12.

[9] Chae, S., Kwon, S., & Lee, D. (2018). "Predicting infectious disease using deep learning and big data.", *International journal of environmental research and public health*, 15(8), 1596.

[10] Twitter, *Twitter*. https://twitter.com/ [accessed Nov. 20, 2021].

[11] Twitter Developer Platform,https://developer.twitter.com/en [accessed Nov. 20, 2021].

[12] Twitter4J - A Java library for the Twitter API, https://twitter4j.org/en/index.html [accessed Nov. 20, 2021].

[13] Kulkarni, A., & Shivananda, A. (2021). *Converting text to features. In Natural language processing recipes* (pp. 63-106). Apress, Berkeley, CA.

[14] Al-Khurayji, R., & Sameh, A. (2017). "An effective arabic text classification approach based on kernel naive bayes classifier.", *Int J Artif Intell Appl*, 8(6), 01-10.

[15] Tsoumakas, G., & Katakis, I. (2007). "Multi-label classification: An overview.", *IJDWM*, 3(3), 1-13.

[16] Varma, M. K. S., Rao, N. K. K., Raju, K. K., & Varma, G. P. S. (2016, February). "Pixel-based classification using support vector machine classifier.", *In 2016 IACC, IEEE*. (pp. 51-55).

[17] Wang, C., Shu, Q., Wang, X., Guo, B., Liu, P., & Li, Q. (2019). "A random forest classifier based on pixel comparison features for urban LiDAR data.", *ISPRS journal of photogrammetry and remote sensing*, 148, 75-86.