

Forecasting epidemic diseases with Arabic Twitter data and WHO reports using machine learning techniques

Qanita Bani Baker, Farah Shatnawi, Saif Rawashdeh

Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan

Article Info

Article history:

Received Jun 8, 2021

Revised Dec 1, 2021

Accepted Feb 28, 2022

Keywords:

Arabic language

Epidemic

Machine learning

Twitter

WHO reports

ABSTRACT

Twitter is one of the essential social media tools used by many people because they express their views, daily problems, and what they suffer from the health aspects. On Twitter, we can detect and track the spread of the most serious diseases like flu; by analyzing people's tweets and collecting reports from health organizations. In this paper, the data from Twitter was collected in the Arabic language related to the spread of influenza using many Arabic keywords. Then, we applied several machine learning algorithms, which are random forest, multinomial naïve bayes, decision tree, and voting classifier. We also found the correlation between the collected tweets and the reports collected from the World Health Organization (WHO) website according to three experiments. These experiments are: i) between the tweets and reports based on the 13 countries regardless of the time, ii) between the tweets and reports based on the Arab regions that depend on these countries' dialects irrespective of the time, iii) between all tweets and all reports based on the week number. The results from these experiments show that there is a strong correlation between the tweets and the reports, which means that the tweets and the WHO reports can together detect the flu outbreaks in the Arab world.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Qanita Bani Baker

Department of Computer Science, Jordan University of Science and Technology

Irbid, Jordan

Email: qmbanibaker@just.edu.jo

1. INTRODUCTION

Twitter is one of the most popular blogs globally, encouraging millions of people to use it and attracting the interest of many researchers [1]. In 2017, there were more than 100 million users on Twitter and more than 5 million tweets per day, with 140 characters for each tweet [1], [2]. For this reason, users express their views on a topic and their daily problems, as well as their fears and their health and psychological conditions [1]-[3]. The spread of infectious diseases is one of the most dangerous events that threatens many people's lives and leads to their death. These diseases are SARS, MERS, Ebola, avian influenza A (H7N9), and seasonal influenza [2], [4]-[6]. Infectious diseases generally have symptoms and indications that indicate that they are infected, such as fever, diarrhea, fatigue, muscle aches, and coughing [7]. One of the deadly events known as the Black Death, which occurred between 1346 and 1353, led to many people's deaths because of the spread of those diseases. About 75 percent of the population of Asia and Europe died. In Europe, 25 million people died [8].

Influenza is a serious viral infection that attacks the respiratory system, including the lungs, nose, and throat. Sometimes, it can lead to serious complications, which in turn can cause some deaths [2], [9], [10]. The times of its spread are in all seasons of the year, especially in the winter. The reason for this is the ease of spread between people through the air, and the infection of a person coughs a lot [9]. All people are vulnerable,

but the most susceptible people are; i) children under five years old, ii) pregnant women, iii) people who have reached the age of one year. Many symptoms of this disease, like headaches, continuous cough, sore throat, and aching muscles [10].

There are several ways in which health agencies know and control the spread of influenza. These methods are the use of social media available on the Internet, the number of visits to the hospital, and finally, the centers for disease control and prevention (CDC) [1], [11]-[13]. Monitoring the disease's spread is very important to reduce and prevent future occurrences [11]-[13]. Social media is the most accurate and fast way to detect early and monitor the disease by collecting data related to the disease from Twitter and Facebook. It provides an early week of other ways to know whether the disease spread or not before it officially appeared through CDC reports [12]-[14]. In different ways, it is collected through visits to clinics, hospitals, and CDC reports to take medical reports and analyze them related to the patients. Despite this method's accuracy, it takes a long time, from about a week to two weeks, to determine whether it is infected or not, and provide it to researchers and organizations to tell people to take precautions [11], [13], [14].

Social media data is enormous, unstructured, not cleaned and meaningful is unknown [15]. To clean the data and build a model to predict the unknown class for each tweet, data mining techniques and preprocessing steps must be used. These steps are tokenization, stemming, stop word removal, and generating term frequency-inverse document frequency (TF-IDF) [16]. The techniques are decision tree, Naïve bayes, K-nearest neighbor, support vector machine (SVM), and others [17]. The contribution of this paper is as; i) studying the influenza disease outbreaks during a specific period in the Arab World. This is done by collecting Arabic tweets and then classifying them as valid or not, ii) building four machine learning algorithms to predict whether tweets are valid or not that are random forest (RF), multinomial naïve bayes (MNB), decision tree (DT), and voting classifier, iii) finding a correlation between the World Health Organization (WHO) reports and the tweets in the Arabic language.

This study is an expanded version of our earlier paper [18], which focused on collecting Arabic tweets related to the flu and utilizing four machine learning techniques to analyze sentiment (valid or invalid tweet) using naive bayes, support vector machines, decision trees, and K-nearest neighbor. The remainder of this paper is organized as: section 2 summarizes some of the past research that has been done in relation to this research. Section 3 demonstrates the system's research process. The results and discussions of this study are described in section 4. The conclusion and suggested future work for this paper are presented in section 5.

2. RELATED WORKS

Several researchers studied infection diseases using machine learning algorithms to predict outbreaks of these diseases, such as our previous paper [18], who used machine learning algorithms to apply sentiment analysis processes on Arabic tweets (valid or invalid). In addition to collecting data from various social media, the researchers collected medical reports from medical resources like the WHO site, centers for disease control (CDC), or hospitals reports.

Alkouz *et al.* [1] used Twitter to collect Arabic tweets as well as the number of flu-related hospital visits in the United Arab Emirates (UAE). The tweets are then classified as reporting, non-self-reporting, or non-reporting. They are also employing a machine learning tool called the linear regression algorithm to identify a link between tweets and the number of hospital visits. The findings demonstrate that there is a strong link between the number of hospital visits and the number of tweets. Lee *et al.* [2] created a model that forecasts flu activity in the future using a multilayer perceptron with backpropagation. Then, for better flu epidemic prediction, they are combining social media data with CDC findings. The data was gathered from social media sites like Twitter and filtered using preprocessing methods, but the CDC reports were gathered from medical clinics. They discovered that combining social media sites with the CDC resulted in a more accurate and reliable prediction of flu activity. Santos *et al.* [3] used machine learning algorithms for two purposes: identifying tweets relating to flu-like sickness or symptoms, and analyzing health-surveillance data from the flu Net project. The Naive Bayes classifier and multiple linear regression models are two of these techniques. To estimate the incidence rates of influenza-like sickness in Portugal, they gathered data from two sources: Twitter data and a search engine query. They discovered that the Naive Bayes has an accuracy of 0.78 and an F-measure of 0.83.

Kim *et al.* [9] collected reports from the CDC in Korea country as a disease outbreak reference, which is called Korea Centers for Disease Control and Prevention (KCDC). Then, they studied the tweets on the Hangeul Twitter to understand the anxiety caused by the flu outbreak in Korea. Finally, they have developed regression models to predict influenza pestilences and track the flu that happens in the real world. They have shown that the correlation coefficients between tweets and reports are highly correlated. Alessa *et al.* [14] used social media data, especially from Twitter, related to the influenza outbreaks, to detect the spreading flu. Social media is faster than the reports from the CDC to track and detect flu outbreaks from 7 to 10 days. Then, they

use techniques to predict the label of the tweets. These techniques are mechanistic models, text mining, graph data mining, machine learning, math/statistical models, and topic models. The results show that using social media data can support early warnings about flu outbreaks. Allen *et al.* [19] detected and monitored the flu outbreaks using tweets from Twitter and formal reports from 30 states in the USA. The tweets are collected by the geographic information system (GIS). Then, they use the support SVM to predict if a tweet is valid or not. Finally, they found a correlation between the tweets and the reports. The results show that the correlation between them is strong.

Aslam *et al.* [20] collected 159,802 tweets from Twitter that related to the flu outbreaks and reports from San Diego country. Then, they find the correlation between the tweets and the reports using liquidation methods based on the types of tweets. The types are non-retweets, tweets with a URL, retweets, and tweets without a URL. In addition, they use SVM to identify tweets, whether valid or invalid. The results show that the correlation between them is a strong correlation, which is 0.93. Aramaki *et al.* [21] used crawling methods to collect tweets from Twitter. Then, they used several machine learning algorithms to extract tweets that were related to the influenza disease outbreaks. These algorithms are Random Forest, Bagging, decision tree, AdaBoost, Naïve Bayes, logistic regression, Nearest Neighbor, SVM with RBF kernel, and SVM with a polynomial kernel. They have also divided the tweets into positive and negative categories. Finally, they analyzed the accuracy and time of multiple machine learning systems. When compared to other algorithms, the best one is SVM with a polynomial kernel, which has an accuracy of 0.756. Culotta *et al.* [22] gathered a large number of tweets from Twitter about two diseases: influenza rates and alcohol sales, respectively. Then they compared them to alcohol sales data from the US Centers for Disease Control and Prevention and the US Census Bureau. He also employed a classifier based on logistic regression that is a bag-of-words document classifier. Finally, they classified the tweets using logistic regression, SVM, and decision trees. When compared to other algorithms, the SVM has a higher accuracy of 83.98.

Culotta *et al.* [23] used a combination of Twitter tweets and CDC records to detect influenza epidemics. They then evaluated regression models to see if there was a link between the tweets and the CDC data. There are two types of regression models: basic linear regression and multiple linear regression. Over a half-million records were collected for more than two months. When comparing simple linear regression to multiple linear regression, the best correlation was 0.78 with the CDC. Santillana *et al.* [24] collected five datasets from several resources: hospital visit records, Google trends, Twitter, and FluNearYou. Then, they selected an ensemble approach that combined the SVM with radial basis function and stacked linear regression. The results show that using this approach gave better accuracy than using SVM or stacked linear regression. Alshammar *et al.* [25] collected data from Twitter to explore self-reported flu cases. They used machine learning algorithms (SVM and random forest (RF)) to detect these cases. The results are compared based on the un-stemmed and stemmed data. In un-stemmed, the results are 0.91 for SVM and 0.90 for RF. While in stemmed, the results are 0.92 for SVM and 0.89 for RF. These results indicate that machine-learning algorithms are effective at detecting flu cases. Buczak *et al.* [26] collected seasonal influenza from fifty states and four regions of the USA. The collected time was between the periods of December 2000 and April 2013. Then, they applied machine learning algorithms and compared them based on PPV, NP, sensitivity, specificity, F0.5, and F3. These algorithms are decision tree, SVM, and RF. The best algorithm for predicting flu cases is RF in terms of specificity, with 0.979 compared with others.

Cheng *et al.* [27] collected the influenza dataset from three resources in Taiwan: The real-time outbreak and disease surveillance system in emergency departments, the National Health Insurance database in outpatient departments, and records of patients from the National Notifiable Disease Surveillance System. The size of each resource is as; 61,076, 124,764, 11,754 records from 22 countries between January 1, 2008, and December 16, 2019, respectively. Then, they used several machine learning algorithms to predict real-time influenza: random forest, support vector regression, autoregressive integrated moving average (ARIMA), extreme gradient boosting, and stacking classifier. The results show that these algorithms accurately predicted real-time influenza-like illness, leading to the medical decision correctly. Dhaka *et al.* [28] collected datasets for two diseases: Dengue (between 2013 and 2017) and Chikungunya (between 2013 and 2016) from an Indian medical website (<https://data.gov.in/>). Then, they implemented an epidemic alert system using four machine learning algorithms with regression types (i.e., random forest, decision, support vector machine, and multiple linear regression) trees to predict the next target for these epidemics. They have shown that multiple linear regression gave the best results in dengue epidemic prediction, while the SVM is the best algorithm for predicting the Chikungunya disease. Wadhwa *et al.* [29] collected datasets from the WHO website that related to COVID-19 in India between 15/02/2020 and 09/05/2020. This dataset contains many parameters: the total number of active cases, the total number of deaths, and the total number of recovered cases all over India. They conducted a data wise analysis on the dataset to predict the number of active cases, deaths, and recovery using different parameters: daily recovery, daily deaths, and increase the rate of covid-19 cases. The results showed that the number of cases, whether active, deaths, or recovered cases, had increased. While the aforementioned cases would be reduced if the government decided to implement the lockdown across India.

3. METHOD

Figure 1 shows the proposed system architecture that works on influenza disease outbreaks through tweets in the Arabic language and collection reports from the WHO website [30]. The tweets are divided into valid, which means the tweets refer to the influenza tweets, and invalid, which means that the tweets are referred to the tweets not related to influenza. The data from tweets are collected based on the location and the time of each tweet. We then find the correlation between the tweets after applying the preprocessing techniques in many tools and the reports from this site. Finally, we used data mining algorithms in the Python tool to build machine learning models that facilitate the prediction process. These models are multinomial naïve bayes (MNB), random forest (RF), decision tree (DT), and voting classifier. The technique used to divide the data into training and testing is hold-out with different sizes (0.1, 0.2, 0.3, 0.4, and 0.5).

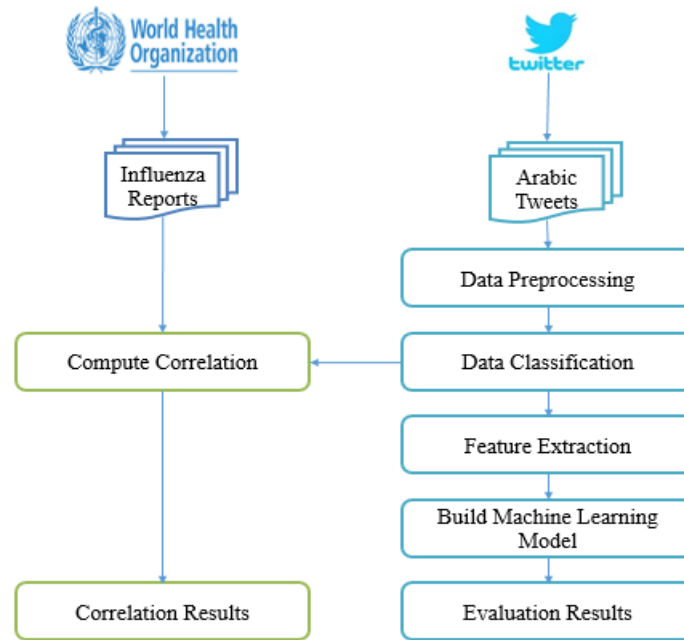


Figure 1. The architecture of the proposed work

3.1. Data and reports collection

In this subsection, the data collection from Twitter and the WHO website is discussed in detail as a following: i) data collection: we collected the data from Twitter by collecting Arabic language tweets based on the nineteen Arab countries' location shown in the following subsection and the time in many steps. The time of collected tweets is from 25/02/2019 to 21/04/2019 or from week 9 to week 16 of 2019; the time is indicated when Twitter tweeted. Firstly, we created a Twitter account to be able to access Twitter. Secondly, The Twitter API is a website [31], which helps us get the keys for data collection. After adding in the Java code, these keys and the twitter4J library [32] allow us to collect the data. In this study, we have collected useful data that are 60,049 out of 166,480 Arabic tweets along with their locations and time and stored in a CSV file format. The data are related to the influenza disease by using 28 keywords like ("سعال cough", "حمى fever", "التهاب الحلق sore throat"). After that, we classified the tweets into two categories that are valid and invalid. Finally, we divide the data into eight weeks related to reports from the WHO website that the data is collected weekly; ii) reports collection: we collected the reports related to the influenza outbreaks from the WHO website through many steps from 25/02/2019 to 21/04/2019 or from week 9 to week 16 of 2019. The first step is choosing the country in which you want to collect its report. Then, select the year and the week number, and click display reports. After this, it should choose the format file in which you want to save the reports. We collected the reports of nineteen Arab countries from week 9 to week 16, which are KSA, Jordan, Bahrain, Tunisia, Iraq, Oman, Palestine, Qatar, Kuwait, Lebanon, Egypt, Morocco, Mauritania, Yemen, UAE, Sudan, Syria, Libya, and Algeria. Some countries do not have any reports, such as Yemen, UAE, Sudan, Syria, Libya, and Algeria. After collecting the data and reports, we find the correlation value between them for thirteen Arab countries using the Linear Regression algorithm. These values need to find it to know if the tweets support and provide the reports to detect the flu outbreaks, or not.

3.2. Data categories

We manually classified the Arabic dataset into two categories: valid, and invalid. Valid tweets indicate that the tweets related to influenza, which are 11,560 tweets. At the same time, the invalid tweets indicate that the tweets are not related to influenza, which is 48,489 tweets. Some Arabic and English translation examples are shown in Table 1.

Table 1. Arabic and English translation examples for the dataset

Data Category	Valid Category	Invalid Category
Tweet Arabic example	الف سلامه عليك ياقلبي انا جسمي مكسر كل عظم بجسمي يعورني وصداع وزكام بس ما فيه حراره.	في وصف الغيرة ستشعر وكأن حرارة هذا العالم عالقة في عينيك.
English translation	Thank God for your safety. My body is exhausted and every bone in my body hurts and I have headaches and colds, but there is no temperature.	In describing the jealousy, you will feel the heat of this world stuck in your eyes.

In the valid example, the words (“صداع وزكام”, " headaches and colds ") indicate that the person is suffered from flu disease. While in the invalid example, the words (“حرارة”, " heat ") indicate that the person is not suffered from flu disease.

3.3. Data preprocessing

After collecting the data, it is not clean and contains a lot of words that are not useful. In this paper, we are applied several preprocessing techniques in three steps. The first step is called One-line Tweet Gathering, which means that we collected each tweet in one line to facilitate the classification process and make the tweet meaningful using code written in Java Programming Language. Twitter's tweets are not in one line and return just 140 characters. So, we need to make each tweet's content in one line and fix any incomplete tweet. The second step is called the duplication removing process, which means that we do not need the duplicated tweets in the classification process by classifying the unduplicated tweets. Because several users post the same tweets or retweets of the tweets, so we removed them using the Excel tool. The final step is called filtering process, which aims to delete English letters, Twitter's users mention (@), punctuation marks, numbers, and emotions. This is done by using code written in the Python Programming Language [27]. The Arabic dataset is now understandable, meaningful, and ready to classify into valid and invalid. So, this processed dataset is stored in an Excel file to facilitate the classification process. We then extracted essential features from the classified dataset, as shown in the features extraction subsection.

3.4. Features extraction

We identify essential features from the processed dataset after preprocessing in order to feed them into machine learning techniques. These algorithms can't handle text as is, they have to transform it to numbers using a variety of methods. We employed a CountVectorizer approach in Python in this work, which counts each tweet's unique word in the dataset. There are 142,126 features in all that use this approach. These algorithms can now deal with text-based on word count in each tweet [33].

3.5. Machine learning algorithms

After applying the preprocessing steps, many machine learning algorithms are used to build the model, facilitating the prediction process. These algorithms are implemented in the Python tool, which are random forest (RF), multinomial naïve bayes (MNB), decision tree (DT), and voting classifiers. We will discuss in detail these algorithms as the following:

Random forest is a machine learning approach in which each object in the dataset has a class label. The RF is an ensemble classifier, which combines more than decision trees to improve overall accuracy [34]. The advantages of RF include the ability to handle huge dimensionality of data and the fact that it is not vulnerable to overfitting as the decision tree approach is [34]. In this paper, the parameters in Python are RF algorithm: **RandomForestClassifier()**, the **criterion** is *Gini index*, the **min_samples_split** is equal to 2, **max_depth** is equal to *None*, the **n_estimators** is equal to 100, and **random_state** is equal to *None*.

Multinomial naïve bayes is a machine learning method that is supervised. It's used for the dataset's classification and prediction. A Bayes assumption is used in this classifier [35]-[37]. This method is concerned with discrete values rather than numerical values, such as sports, social or economic categories, and valid or invalid categories, as shown in equation 1 [38], where the P(H) is "hypothesis probability of H," P(X) is "evidence probability," P(X|H) is "the X's probability on H is true," and P(H|X) is "the H's on X," as shown in (1):

$$P(H|X)=P(X|H)*P(H)/P(X) \quad (1)$$

In this paper, we implemented the MNB in Python tool, which the parameters are MNB algorithm: **naive_bayes.MultinomialNB()**, **fit_prior** with *True*, **alpha** with value *1.0*, and **class_prior** (number of class) with *None*.

Decision tree is a machine learning algorithm with a supervised type. It's used for the dataset's categorization and prediction. This approach creates a dataset tree with leaf nodes representing class values and internal nodes and root nodes representing characteristics [39]. Using the Gini index and information gain, the decision tree selects the best attribute as the tree's root. J48, logistic model trees, reduced error pruning trees, and alternating decision trees are just a few examples [40]. In this paper, we implemented the DT in Python tool, which the parameters are DT algorithm: **DecisionTreeClassifier()**, **splitter** is *best*, the **min_samples_split** is equal to 2, the **criterion** is *Gini index*, **min_samples_leaf** is equal to 1, **max_depth** is equal to *None*, and **random_state** is equal to *None*.

Voting classifier is a type of ensemble classifier that is used to improve classification and prediction accuracy [41]. It uses the same type of base classifiers or a combination of different types, such as MNB, RF, or NB, and KNN at the same time [41]. The essential premise of the Voting classifier is that each base classifier predicts an unknown dataset, and then each classifier's prediction is combined using the average approach or majority voting based on the class label type [41]. Using the average for the continuous label, the discrete label received the majority vote [41]. In this paper, we implemented the four classifiers that combined them in the Python tool in the Voting classifier. The parameters of this classifier are as a following: Voting algorithm: **VotingClassifier()**, the **voting type** is *hard*, and **base classifiers** are *RF, MNB, and DT*.

3.6. Machine learning algorithms results

To ensure and check results after applying these algorithms in Python, we re-run the code five times and calculate the average results based on accuracy, precision, recall, and F1-score as shown in Figures 2, 3, 4, and 5, respectively. In Figure 2, the best algorithm that gave 97.53% accuracy is the VC algorithm in a test size of 0.5. For Figure 3, all algorithms except MNB gave the highest precision values of 97.8% in test size set to 0.5. But in Figure 4, the best algorithm that gave 97.8% of recall is the DT algorithm in test size of 0.5. Finally, in Figure 5, the best algorithms that gave 97.4% of the F1-score is the RF algorithm in a test size of 0.5. These results indicate that the RF, DT, and VC algorithms gave the best performance in the prediction process on the influenza dataset used in this paper. While the MNB gave low results when applied to this dataset. The results of the accuracy, precision, recall and F1-score of the aforementioned machine learning algorithms prove that we can forecast flu outbreaks in the Arab world in the future.

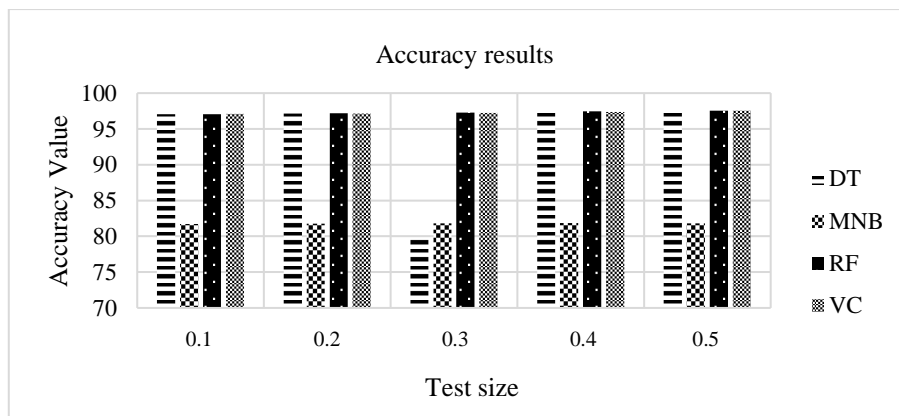


Figure 2. Accuracy values of machine learning algorithms

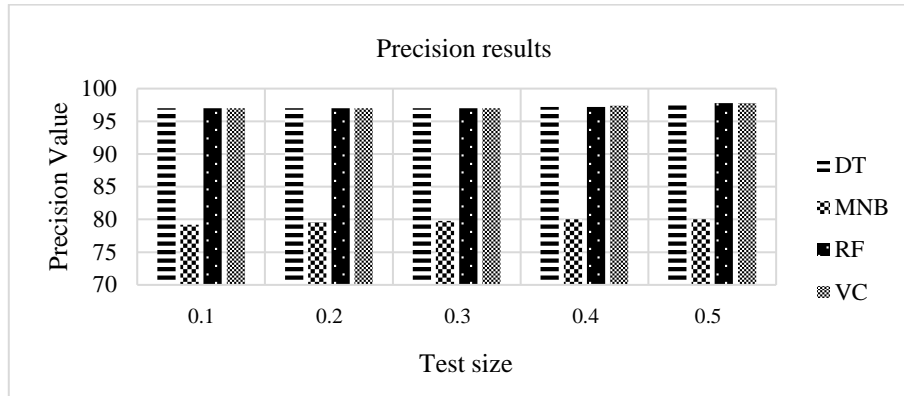


Figure 3. Precision values of machine learning algorithms

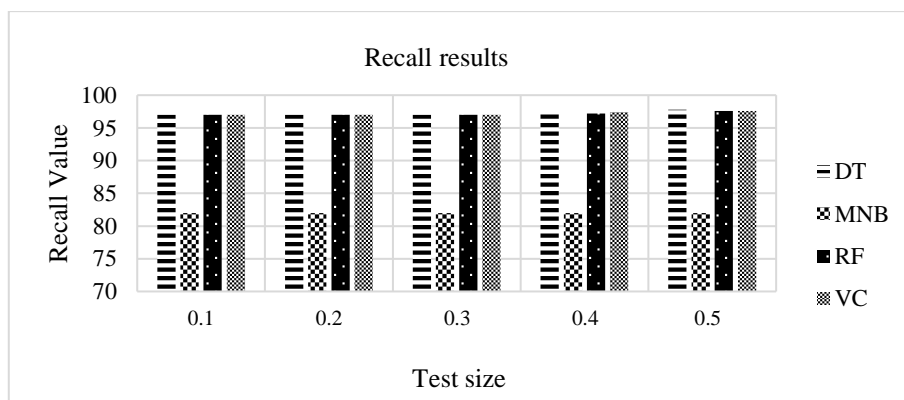


Figure 4. Recall values of machine learning algorithms

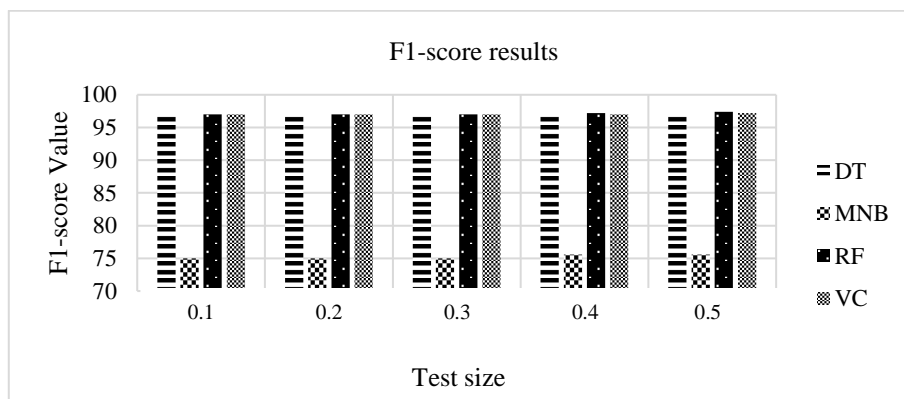


Figure 5. F1-score values of machine learning algorithms

4. RESULTS AND DISCUSSION

In this section, we explained the Correlation value results between the tweets from Twitter and the reports from the WHO website for nineteen Arab countries, which are mentioned in the previous section using the Linear Regression algorithm. However, six countries do not have reports; the number of countries is thirteen in tweets and reports. The time of both datasets whether tweets or reports are collected between 25/02/2019 and 21/04/2019, on the other hand between week 9 of 2019 and week 16 of 2019. We conducted three experiments on both datasets based on calculating the correlation between: i) tweets and reports based on the 13 countries, regardless of the time; ii) tweets and reports based on the Arab regions that depend on these countries' dialects irrespective of the time; iii) all tweets and all reports are based on the week's number.

4.1. First experiment

This experiment is conducted on the thirteen Arab countries and finds the correlation between the tweets and reports in two cases. The two cases are; i) positive cases or valid cases, which means that the tweet or report is related to influenza; ii) all cases include all the people who visit the hospital and valid tweets. The correlation value of linear regression is greater than zero, zero, and less than zero. If the value is greater than 0, the correlation is positive; the number of tweets increases, and the number of reports increases (both variables change in the same direction). If the value is less than 0, the correlation is negative; the number of tweets decreases, and the number of reports increases (both variables change in the opposite direction). At the same time, no correlation if the value is equal to 0. These countries are KSA, Jordan, Bahrain, Tunisia, Iraq, Oman, Palestine, Qatar, Kuwait, Lebanon, Egypt, Morocco, Mauritania. Figure 6 shows the correlation values, which is represented in the y-axis between the tweets and reports that calculated on the thirteen Arab countries, which is represented in the x-axis. The highest correlation in the first case is in Oman with a positive correlation that is 0.46. The highest correlation in the second case in Mauritania with a positive correlation that is 0.7329.

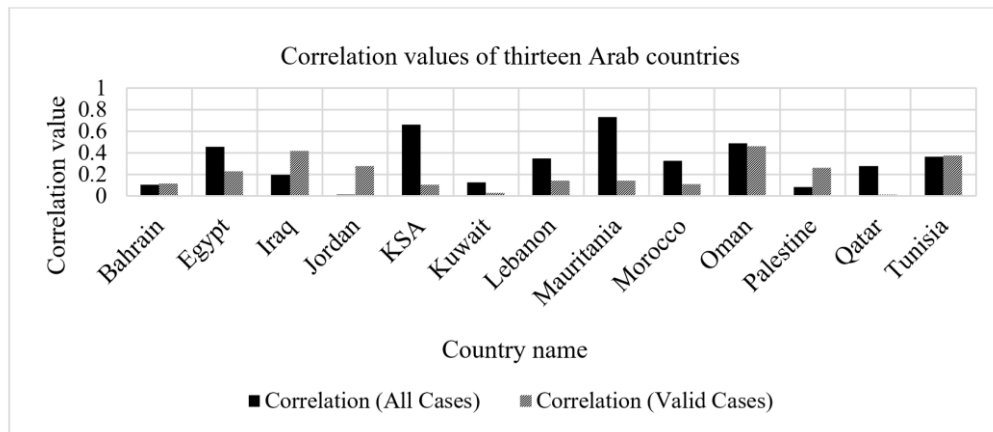


Figure 6. Correlation values of thirteen countries

4.2. Second experiment

This experiment is conducted on the five Arab regions and finds the correlation between the tweets and reports in two cases. These regions are Levant, Iraq, Nile countries, Arab Gulf, and Maghreb countries. The Levant region includes Jordan, Palestine, Lebanon, and Syria. The Iraq region has only Iraq country. The Arab Gulf region includes KSA, Bahrain, Oman, Qatar, Kuwait, Yemen, and UAE. The Maghreb countries region includes Tunisia, Morocco, Mauritania, Libya, and Algeria. The Nile countries region has Egypt and Sudan. Figure 7 shows the correlation values, which is represented in the y-axis between the tweets and reports that calculated on the 5 Arab regions, which is represented in the x-axis. The highest correlation in the first case in the Levant region with a positive correlation is 0.6626. The highest correlation in the second case is in Nile countries with a positive correlation that is 0.5356.

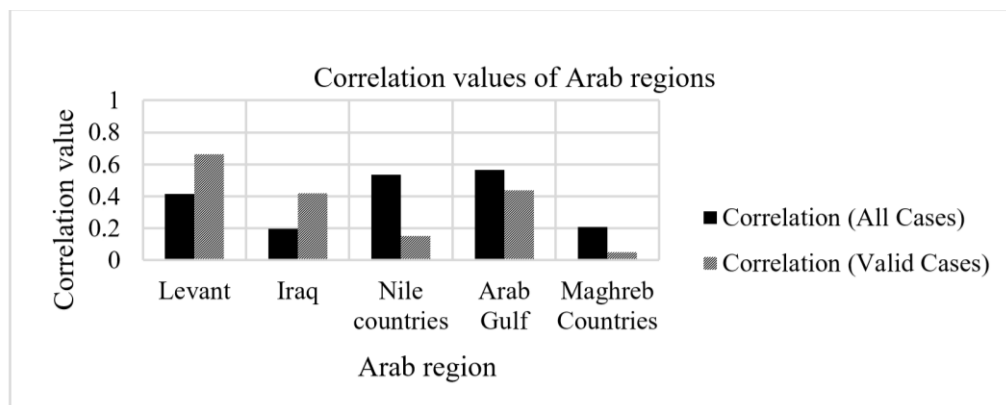


Figure 7. Correlation values of Arab regions

4.3. Third experiment

This experiment is conducted on tweets and reports in terms of summation of all tweets and all reports for each week number and finding the correlation between the tweets and reports in two cases using the linear regression algorithm. The week number is from 9 to 16 of the year 2019, on the other hand from 25/02/2019 to 21/04/2019. Figure 8 shows the correlation values between the all tweets and all reports that are calculated. The correlation in the first case is a positive correlation that is 0.534. At the same time, the correlation in the second case is positive that is 0.441.

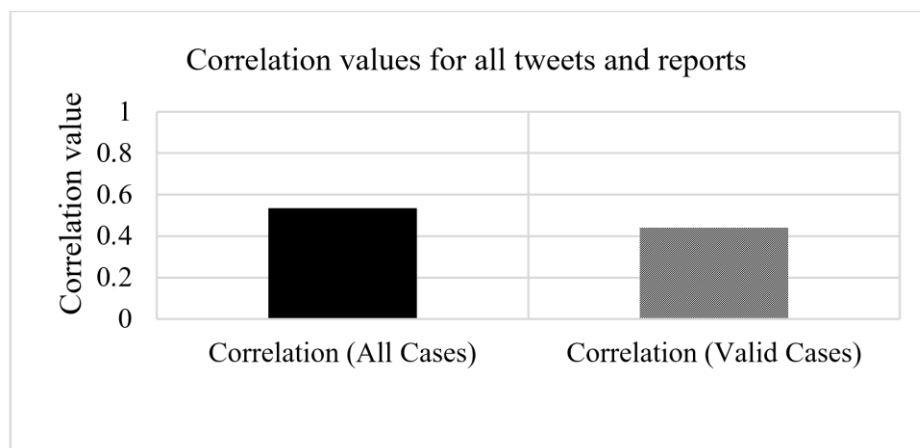


Figure 8. Correlation values for all tweets and reports

5. CONCLUSION

This paper discusses the relationship between tweets from Twitter and medical reports from the WHO website using correlation and linear regression. This relationship between the tweets and reports has been calculated in three experiments i) between the tweets and reports based on the 13 countries regardless of the time; ii) between the tweets and reports based on the Arab regions that depend on these countries' dialects irrespective of the time; iii) between all tweets and all reports based on the week number. So, to see if there is any change in the different countries or regions in the Arab countries. The results from these experiments show a correlation between the tweets and the reports, which means that the tweets and the reports can together detect the flu outbreaks in the Arab world. The future work of this paper will be: i) collecting reports from hospitals related to flu outbreaks; ii) increasing the dataset size, including all the Arab countries equally in terms of the number of tweets and reports; iii) using more machine learning and deep learning algorithms; iv) applying the natural language processing (NLP) techniques to the data.

REFERENCES




- [1] B. Alkouz and Z. A. Aghbari, "Analysis and prediction of influenza in the UAE based on Arabic tweets," *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, 2018, pp. 61-66, doi: 10.1109/ICBDA.2018.8367652.
- [2] K. Lee, A. Agrawal and A. Choudhary, "Forecasting Influenza Levels Using Real-Time Social Media Streams," *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017, pp. 409-414, doi: 10.1109/ICHI.2017.68.
- [3] J. Santos and S. Matos, "Analysing Twitter and web queries for flu trend prediction", *Theoretical Biology and Medical Modelling*, vol. 11, no. 1, pp. 1-11, 2014, doi: 10.1186/1742-4682-11-S1-S6.
- [4] R. Bernard, G. Bowsher, C. Milner, P. Boyle, P. Patel and R. Sullivan, "Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks," *Journal of Public Health*, vol. 26, no. 5, pp. 509-514, 2018, doi: 10.1007/s10389-018-0899-3.
- [5] I. Fung, K. Fu, Y. Ying, B. Schaible, Y. Hao, C. Chan and Z. Tse, "Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks," *Infectious Diseases of Poverty*, vol. 2, oo. 1, pp. 1-12, 2013, doi: 10.1186/2049-9957-2-31.
- [6] A. O. Williams, J. Warren, L. Kurlander, and M. Suaray, "Critical Communications: A Retrospective Look at the Use of Social Media among American Sierra Leoneans during the Ebola Outbreak," *The Journal of Social Media in Society*, vol. 7, no. 1, pp. 366-380, 2018.
- [7] S. Lim, C. Tucker and S. Kumara, "An unsupervised machine learning model for discovering latent infectious diseases using social media data," *Journal of Biomedical Informatics*, vol. 66, pp. 82-94, 2017, doi: 10.1016/j.jbi.2016.12.007.
- [8] W. Ahmed, P. Bath, L. Sbaffi and G. Demartini, "Moral Panic through the Lens of Twitter," *Proceedings of the 9th International Conference on Social Media and Society*, pp. 217-221, 2018, doi: 10.1145/3217804.3217915.
- [9] E. Kim, J. Seok, J. Oh, H. Lee and K. Kim, "Use of Hangeul Twitter to Track and Predict Human Influenza Infection," *PLoS ONE*, vol. 8, no. 7, p. e69305, 2013, doi: 10.1371/journal.pone.0069305.

- [10] A. Monto, S. Gravenstein, M. Elliott, M. Colopy and J. Schweinle, "Clinical Signs and Symptoms Predicting Influenza Infection," *Archives of Internal Medicine*, vol. 160, no. 21, pp. 3243-3247, 2000, doi: 10.1001/archinte.160.21.3243
- [11] K. Lee, A. Agrawal and A. Choudhary, "Real-time disease surveillance using Twitter data," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, pp. 1474-1477, 2013, doi: 10.1145/2487575.2487709.
- [12] V. Jain and S. Kumar, "An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter," *Procedia Computer Science*, vol. 70, pp. 801-807, 2015, doi: 10.1016/j.procs.2015.10.120.
- [13] C. St Louis and G. Zorlu, "Can Twitter predict disease outbreaks?," *BMJ*, vol. 344, no. 172, pp. e2353-e2353, 2012, doi: 10.1136/bmj.e2353.
- [14] A. Alessa and M. Faezipour, "A review of influenza detection and prediction through social networking sites," *Theoretical Biology and Medical Modelling*, vol. 15, no. 1, pp. 1-27, 2018, doi: 10.1186/s12976-017-0074-5.
- [15] A. Holzinger and I. Jurisica, "Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions," *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, Heidelberg, Berlin, pp. 1-18, 2014, doi: 10.1007/978-3-662-43968-5_1.
- [16] R. Saravanan and M. Rajesh Babu, "Enhanced text mining approach based on ontology for clustering research project selection," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-11, 2017, doi: 10.1007/s12652-017-0637-7.
- [17] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *arXiv preprint arXiv:1707.02919*, 2017.
- [18] Q. B. Baker, F. Shatnawi, S. Rawashdeh, M. Al-Smadi, and Y. Jararweh, "Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets," *Journal of Universal Computer Science*, vol. 26, no. 1, pp. 50-70, 2020.
- [19] C. Allen, M. Tsou, A. Aslam, A. Nagel and J. Gawron, "Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza," *PLOS ONE*, vol. 11, no. 7, pp. 1-10, 2016, doi: 10.1371/journal.pone.0157734.
- [20] A. A. Aslam *et al.*, "The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance", *Journal of Medical Internet Research*, vol. 16, no. 11, p. e250, 2014, doi: 10.2196/jmir.3532.
- [21] E. Aramaki, S. Maskawa, and M. Morita, "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1568-1576, 2011.
- [22] A. Culotta, "Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 217-238, 2012, doi: 10.1007/s10579-012-9185-0.
- [23] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pp. 115-122, 2010, doi: 10.1145/1964858.1964874.
- [24] M. Santillana, A. Nguyen, M. Dredze, M. Paul, E. Nsoesie and J. Brownstein, "Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance," *PLOS Computational Biology*, vol. 11, no. 10, p. e1004513, 2015, doi: 10.1371/journal.pcbi.1004513.
- [25] S. Alshammari and R. Nielsen, "Less is More: With a 280-character limit, Twitter Provides a Valuable Source for Detecting Self-reported Flu Cases," *Proceedings of the 2018 International Conference on Computing and Big Data - ICCBD '18*, pp. 1-6, 2018, doi: 10.1145/3277104.3277105.
- [26] A. Buczak, B. Baugher, E. Guven, L. Moniz, S. Babin and J. Chretien, "Prediction of Peaks of Seasonal Influenza in Military Health-Care Data," *Biomedical Engineering and Computational Biology*, vol. 72, pp. 15-26, 2016, doi: 10.4137/BECB.S36277.
- [27] H. Cheng *et al.*, "Applying Machine Learning Models with An Ensemble Approach for Accurate Real-Time Influenza Forecasting in Taiwan: Development and Validation Study," *Journal of Medical Internet Research*, vol. 22, no. 8, p. e15394, 2020, doi: 10.2196/15394.
- [28] A. Dhaka and P. Singh, "Comparative Analysis of Epidemic Alert System using Machine Learning for Dengue and Chikungunya," *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2020, pp. 798-804, doi: 10.1109/Confluence47617.2020.9058048.
- [29] P. Wadhwa, Aishwarya, A. Tripathi, P. Singh, M. Diwakar and N. Kumar, "Predicting the time period of extension of lockdown due to increase in rate of COVID-19 cases in India using machine learning", *Materials Today: Proceedings*, vol. 37, pp. 2617-2622, 2021, doi: 10.1016/j.matpr.2020.08.509.
- [30] Apps.who.int. 2021. WHO FLUMART OUTPUTS. [online] Available at: <<http://apps.who.int/flumart/Default?ReportNo=12>> [Accessed 7 December 2021].
- [31] Use cases, tutorials, & documentation | twitter developer platform, Twitter. [Online]. Available: <https://developer.twitter.com/>. [Accessed: 27-Jan-2022].
- [32] A Java library for the Twitter API, Twitter4J. [Online]. Available: <http://twitter4j.org/en/>. [Accessed: 27-Jan-2022].
- [33] A. Kulkarni and A. Shivananda, "Converting Text to Features," *Natural Language Processing Recipes*, pp. 67-96, 2019, doi: 10.1007/978-1-4842-4267-4_3.
- [34] C. Wang, Q. Shu, X. Wang, B. Guo, P. Liu and Q. Li, "A random forest classifier based on pixel comparison features for urban LiDAR data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 148, pp. 75-86, 2019, doi: 10.1016/j.isprsjprs.2018.12.009.
- [35] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier," *International Journal of Information Engineering and Electronic Business*, vol. 8, no. 4, pp. 54-62, 2016, doi: 10.5815/ijeeb.2016.04.07
- [36] D. S. Vijayarani and M. S. Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 4, no. 4, pp. 816-820, 2015.
- [37] B. Tang, S. Kay and H. He, "Toward Optimal Feature Selection in Naive Bayes for Text Categorization," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2508-2521, 1 Sept. 2016, doi: 10.1109/TKDE.2016.2563436.
- [38] T. A. Shinde and D. J. R. Prasad, "IoT based Animal Health Monitoring with Naive Bayes Classification," *IJETT*, vol. 1, no. 2, 2017.
- [39] R. Pandya and J. Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *International Journal of Computer Applications*, vol. 117, no. 16, pp. 18-21, 2015.
- [40] K. Khosravi *et al.*, "A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran," *Science of The Total Environment*, vol. 627, pp. 744-755, 2018, doi: 10.1016/j.scitotenv.2018.01.266.




- [41] J. Matoušek and D. Tihelka, "Glottal Closure Instant Detection from Speech Signal Using Voting Classifier and Recursive Feature Elimination," *Proceeding of Interspeech*, pp. 2112-2116, 2018, doi: 10.21437/Interspeech.2018-1147.

BIOGRAPHIES OF AUTHORS






Dr. Qanita Bani Baker    is currently an Associate Professor of Computer Science at Jordan University of Science and Technology (JUST). She received her Ph.D. in Computer Science from Utah State University in the USA in 2015. Dr. Baker has worked at JUST since 2015. Her research interests include Data Science, problem optimization, big data, bioinformatics, high-performance computing, evolutionary algorithms, and capacity building. Dr. Baker is currently conducting several studies in optimizing and analyzing biomedical big data and tools. Dr. Baker has co-authored many technical papers in specialized peer-reviewed international journals and conferences. She can be contacted at qmbanibaker@just.edu.jo.



Farah Shatnawi    received Bachelor of Computer Science in 2015 from Al-Balqa' Applied University, Irbid, Jordan, and Master of Computer Science from Jordan University of Science and Technology, Irbid, Jordan in 2020. The Master thesis is about Detecting Humor in English News Headlines: A Comprehensive Study of Pre-Trained Deep Learning Models. From 2019 until now, she is a Research Assistant in the Data Science field at the Computer Science department with many areas, such as Healthcare and Educational areas. She can be contacted at email: ffshatnawi16@cit.just.edu.jo.



Saif Rawashdeh    received Bachelor of Computer Science in 2016 from Jordan University of Science and Technology, Irbid, Jordan, and Master of Computer Science from Jordan University of Science and Technology, Irbid, Jordan in 2019. The master project is about forecasting the flu disease using machine learning models and WHO reports in the Arabic Language. Currently, he is the Research Assistant at the Computer Science department from 2019 until now in Jordan University of Science and Technology in the Data Science field with many areas like Educational and Healthcare areas. He can be contacted at email: sarawashdeh16@cit.just.edu.jo.