Journal of Universal Computer Science, vol. 26, no. 1 (2020), 50-70
submitted: 30/12/18, accepted: 15/11/19, appeared: 28/1/20

Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets

Qanita Bani Baker [1]
(Jordan University of Science and Technology, Irbid, Jordan
qmbanibaker@just.edu.jo)

Farah Shatnawi
(Jordan University of Science and Technology, Irbid, Jordan
ffshatnawi16@cit.just.edu.jo)

Saif Rawashdeh
(Jordan University of Science and Technology, Irbid, Jordan
sarawashdeh16@cit.just.edu.jo)

Mohammad Al-Smadi
(Jordan University of Science and Technology, Irbid, Jordan
maalsmadi9@just.edu.jo)

Yaser Jararweh
(Jordan University of Science and Technology, Irbid, Jordan
yijararweh@just.edu.jo)

Abstract: Opinion mining is an important step towards facilitating information in health data. Several studies have demonstrated the possibility of tracking diseases using public tweets. However, most studies were applied to English language tweets. Influenza is currently one of the world's greatest infectious disease challenges. In this study, a new approach is proposed in order to detect Influenza using machine learning techniques from Arabic tweets in Arab countries. This paper is the first study of epidemic diseases based on Arabic language tweets. In this work, we have collected, labeled, filtered and analyzed the influenza-related tweets written in the Arabic language. Several classifiers were used to measure the quality and the performance of the approach, which are: Naive Bayes, Support Vector Machines, Decision Trees, and K-Nearest Neighbor. The classifiers which achieved the best accuracy results for the three experiments were: Naïve Bayes with 89.06%, and K-Nearest Neighbor with 86.43%, respectively.

Keywords: Twitter, Infectious Diseases, Influenza, Arabic Tweets, Sentiment Analysis, Machine Learning, Data Mining.

Category: L.2, J.3, I.2

---
[1] Corresponding author.

1 Introduction

With the spread of smartphones, it becomes easier to access the internet and social networking sites such as Facebook, Twitter, Snapchat and Instagram. Users can publish on these websites their attitudes, feelings and personal experiences, even what they suffer from either as pain or disease [Santos and Matos, 2014] [Suarez et al., 2018]. Through social media sites, we can analyze people's concerns and worries as well as finding some infectious diseases during a certain period of time in a particular country throughout what people share in their posts or tweets. One of these sites is Twitter, which is one of the most popular social networking sites where people can publish their personal information and even their physical problems like infectious or chronical diseases [Fung et al., 2013].

The spread of infectious diseases is one of the most dangerous problems in the world such as Influenza, SARS, MERS, and Ebola [Santos and Matos, 2014], [Fung et al., 2013], [Quwaider and Jararweh, 2016], and [Bernard et al., 2018]. The infectious disease affects the people who are surrounding the patient or in direct touch with them [Ahmed et al., 2018]. Health researchers work for studying the reasons behind these diseases in order to find a way to discover it at an early stage and limit their spread [Allen et al., 2016], [Ahmed et al., 2018] and [Al-Zinati et al., 2019].

Public health agencies depend on traditional ways to control and monitor the expansion of infectious diseases. This way relies on the laboratory reports and doctor's diagnosis, but it takes a long time to detect if the disease spreads or not. Sometimes the disease discovery by social media could be faster than the medical reports [St Louis and Zorlu, 2012]. Recently, it becomes easier and more popular for people to see where the latest infectious diseases are occurring. This can be easily tracked by looking at the posts and tweets shared by others on their personal accounts [Ye et al., 2016]. Several researchers have collected twitter tweets for sentiment analysis and in different languages [Aramaki et al., 2011], [Smadi and Qawasmeh, 2018], but there are few numbers of research in the Arabic language that target health issue and diseases. Our paper is the first study to investigate if epidemic disease can be detected or not based on Arabic tweets.

Influenza is a dangerous viral infectious disease that sometimes causes death [Lee et al., 2017]. Influenza can transfer in several ways, such as coughing, air, oral saliva, sneezing or even when talking with sufferers [Kim et al., 2013]. Influenza is characterized by a sudden body temperature; sore throat, headache, muscle and joint pain, nausea, cough that usually dry and runny nose. A severe cough can last for two weeks or more. Most patients recover from fever and other symptoms within one week without the need for medical attention. The period between infection and the onset of the disease is known as the incubation period, It lasts about two days WHO. This paper examines Arabic tweets data to detect influenza epidemics. These tweets are analyzed using several data mining techniques. The tweets were manually labeled into ”valid” or ”invalid” by native Arabic speakers. The goal of this study is to discover whether the influenza disease can be detected through the study of Arabic tweets in the geographical regions of the Arab world countries.

The remaining sections of this paper are organized as follows: Section 2 describes the literature review. Section 3 provides insights into the methodology of the proposed system and describes the experimental design. Section 4 presents the experimental results and the discussion of the findings. In Section 5, we conclude the most important findings from the topic under research. Finally, in Section 6, we present the future work and inform the development of further studies.

2 Literature Review

There are many data mining algorithms applied to detect infectious disease outbreaks such as influenza and Ebola by using social media, which is sometimes faster than health agencies like the Centers for Disease Control and Prevention (CDC). In [Santos and Matos, 2014], Santos et al. highlighted the using of tweets from Twitter and queries in a search engine to predict the influenza-like illness incidence rates in Portugal. Then, they used a Naïve Bayes classifier to determine the tweets that are related to flu-like illness or symptoms. They also utilized multiple linear regression models to appreciate the health-surveillance data from the project on flu Net. Similarly, in [Fung et al., 2013], Fung et al. used the Weibo website to gauge the reaction from people in China during two outbreak diseases. These outbreak diseases are the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) outbreak in 2012 and the outbreak of human infection with avian influenza A (H7N9) in 2013. They collected posts from the web-based on sampling criteria where the users in Weibo have more than 10000 followers. Then, they used the keyword detection method based on a specific keyword like avian flu and H7N9 by searching in the millions of posts depends on these keywords. The results show that the people in China reacted with the two outbreaks of diseases in social media.

In [Ahmed et al., 2018], Ahmed et al. reviewed qualitative analysis to analyze the tweets which are related to the kinds of infectious disease outbreaks such as Swine Flu and Ebola. The number of tweets that are collected from the Firehose API (set of tweets that are collected from Twitter by a licensed reseller) is 13,373 tweets and used thematic analysis to analyze these tweets. In [Allen et al., 2016], Allen et al. used geographic information science techniques (GIS) to collect and analyze the tweets from Twitter. These techniques are spatial filtering, populated normalization, and multi-scale analysis. The tweets were related to the flu outbreaks in 30 cities which are the most densely populated cities in America. Then, they used a machine-learning algorithm to classify the tweets if they are related to influenza or not (valid or invalid tweets) based on keywords such as flu and influenza using Support Vector Machine (SVM). To train the SVM classifier, they used 1500 tweets as a sample. In [Ye et al., 2016], Ye et al. developed a web crawler to collect the data from Weibo. This data source is similar to Twitter, but the data were in the Chinese language. They used this data source to explore how infectious diseases spread. Then, they analyzed the dengue fever, which is shown in Weibo messages using spatial analysis, temporal analysis, and spatiotemporal pattern.

In [Lee et al., 2017], Lee et al. applied a model using a multilayer perceptron with backpropagation that estimates the flu activities. They integrated the social media and CDC for an accurate prediction of the flu outbreak. The data from social media like Twitter are collected using Twitter API and filtered by using preprocessing steps. They also used the data from the CDC collected from medical practices. In [Kim et al., 2013], Kim et al. collected reports from Korea Centers for Disease Control and Prevention (KCDC) as a disease outbreaks reference and studied the tweets in the Hangeul twitter to detect anxiety and develop rapid public awareness regarding with influenza outbreaks. To predict influenza pestilences in the real world and to follow the disease activity, they developed the regression models. In [Bernard et al., 2018], Bernard et al. presented how to use different tools from the sector of clandestine intelligence to detect infectious disease outbreak cases such as SARS, MERS, Ebola. These tools are Open Source Intelligence (OSINT) and Signals Intelligence (SIGINT).

In [Aramaki et al., 2011], Aramaki et al. used crawling methods for Twitter tweets collection. They used Support Vector Machine (SVM) to extract the tweets that are related to the influenza disease by searching for the ”influenza” word in each tweet. Also, they classified the tweets to positive and negative tweets. Finally, they compared several machine learning algorithms based on accuracy and time. In [Alessa and Faezipour, 2018], Aramaki et al. reviewed several methods to discover the outbreaks of flu using social media such as Twitter. The methods are graph data mining, text mining, topic models, Machine learning techniques, math/statistical models and mechanistic models. In [Wang et al., 2018], Wang et al. combined the advantages of the Vaccine Adverse Event Reporting System (VAERS) information and social media such as Twitter to determine the possible risks after taking the flu vaccine for every person with the flu. Also, they used SVM (linear), SVM (polynomial kernel), SVM (radial basis kernel), Logistic Regression, Neural Network, and Multi-instance Logistic Regression. After that, they compared them based on five metrics Accuracy (ACC), Precision (PR), Recall (RE), F-score (FS) and Area under the ROC

(AUC). In [Lee et al., 2013], Lee et al. described a system to detect the flu and cancer diseases automatically using spatial (Geographical Analysis), temporal Mining, and text mining for Twitter tweets. This approach is called the real-time flu and cancer surveillance system.

In [Culotta, 2013], Culotta collected over half a billion tweets from Twitter that are related to influenza rates and alcohol sales. They analyzed them versus the U.S Centers for Disease Control and Prevention of influenza and U.S. Census Bureau for alcohol sales. Then, Culotta used a document classifier to filter the messages which are not related to these diseases. This classifier is a bag-of-words document classifier using logistic regression. Finally, Culotta used the logistic regression, SVM and Decision tree to classify the tweets. In [Aslam et al., 2014], Aslam et al. collected 159,802 Twitter tweets that contain flu as a keyword from eleven USA cities. They have used two methods to monitor the correlation between the rates of influenza-like illness and tweets. These methods are liquidation of the tweets based on type such as: "non-retweets, retweets, tweets with a URL, tweets without a URL)". They used machine learning algorithms to classify the tweets into valid or invalid. In [van de Belt et al., 2018], van de Belt et al. used posts of social media and Google trends to detect the methicillin-resistant Staphylococcus aureus (MRSA) outbreaks or not. Social media is easier and faster than the health agencies to detect MRSA outbreaks.

In [Chew and Eysenbach, 2010], Chew et al. suggested and assessed an approach using Twitter in H1N1 that happened in 2009. This approach is a complementary surveillance system. They used a surveillance system to collect more than 2 million tweets based on many keywords such as swine flu. In [Culotta, 2010], Culotta analyzed the tweets that are collected from Twitter to explore the influenza outbreaks. Culotta compared between the different numbers of regression models to find the correlation between CDC statistics with the tweets. The number of tweets that Culotta collected is more than a half-million for more than two months. The best Correlation was 0.78 with CDC for Simple Linear Regression. In [Ahmed et al., 2018], Ahmed et al. collected 214,784 tweets from Twitter based on keywords like 'Flu', 'swine flu', and 'H1N1' to detect two infectious disease outbreaks (Swine Flu and Ebola) during a two-day period in April month. In [Signorini et al., 2011], Signorini et al. used the data embedded in the Twitter stream to do two things. First thing, follow the real activity of H1N1 disease. Second thing, follow the sentiment for people that related to H1N1 or swine flu.

In Table 1, we provide a summary of the studies that study several diseases using data collected from social media platforms. Table 1 shows the accuracy and other measurements for each classifier in each of the previous studies. In Table 1, the term "NM*" means that the information was not mentioned in the paper.

Ref.	Method Name	The accuracy or other measurements
[Santos and Matos, 2014]	1) Naïve Bayes classifier (NB). 2) Multiple linear regression (MLR) model	1) The result of NB is 0.78, 0.83 for precision and f-measure respectively. 2) The Correlation ratio of MLR is 0.89.
[Fung et al., 2013]	keyword detection method	NM*
[Ahmed et al., 2018]	Thematic analysis	NM
[Allen et al., 2016]	Support Vector Machine (SVM)	1) Precision score is 0.67 2) Recall score is 0.949 3) F1 score is 0.786
[Ye et al., 2016]	1) Spatial analysis 2) Temporal analysis 3) Spatiotemporal pattern	NM
[Lee et al., 2017]	Presented the model using multilayer perceptron with Back-propagation	NM
[Kim et al., 2013]	Linear Regression	The Regression coefficient is 2:277.
[Bernard et al., 2018]	1) Open Source Intelligence (OSINT) 2) Signals Intelligence (SIGINT)	NM
[Aramaki et al., 2011]	1) AdaBoost 2) Bagging 3) Decision Tree 4) Logistic Regression 5) Naïve Bayes 6) Nearest Neighbor 7) Random Forest 8) SVM (RBF kernel) 9) SVM (polynomial kernel; d=2)	The SVM with polynomial gave the best f-measure value in 0.756.
[Alessa and Faezipour, 2018]	Data mining, text mining, topic models, Machine learning techniques, math/statistical models and mechanistic models	NM
[Wang et al., 2018]	1) SVM(linear) 2) SVM(polynomial kernel) 3) SVM(radial basis kernel) 4) Logistic Regression 5) Neural Network 6) Multi-instance Logistic Regression	1) The Multi-instance Logistic Regression gave the best accuracy score (0.8054), precision score (0.7871), F1 score (0.6984) and AURoc (0.8902). 2) The SVM with radial gave the best recall score (0.9344).
[Lee et al., 2013]	Using real-time flu and cancer surveillance system	NM
[Culotta, 2013]	1) Logistic regression 2) SVM 3) Decision tree	1) The SVM gave the best results in accuracy (83.98), f1 (90.01), and precision (94.38). 2) The Logistic regression gave the best recall score in 94.89.
[Aslam et al., 2014]	Machine learning classifier	The correlation is 0.93.
[van de Belt et al., 2018]	SO-ZI/AMR system	NM
[Chew and Eysenbach, 2010]	complementary infoveillance	NM
[Culotta, 2010]	1) Simple Linear Regression 2) Multiple Linear Regression	The correlation for Simple is 0.78 and for the Multiple is 0.739.
[Ahmed et al., 2018]	Using data from Twitter	NM
[Signorini et al., 2011]	Data in the Twitter stream	NM

Table 1: shows the accuracy and other measurements for each classifier in each reference number.

3 Methodology

In this section, we explain the used methodology applied in this work. Figure 1 shows the flow chart of the proposed system that is utilized to detect the influenza disease epidemic through collected tweets in the Arabic language. We divided the collected data into two groups: invalid tweets that are not related to influenza and valid tweets that are related to influenza. The collected dataset contains the tweets, locations, and the ground truth for each tweet. We analyzed the tweets that are collected from Twitter based on the location and for several Arabic countries. Also, we applied the preprocessing techniques on the data such as tokenization, filter stop words, n-grams and stemming. Finally, we compared several data mining techniques based on accuracy values. These techniques are Support Vector Machine (SVM), Naïve Bayes (NB), K-nearest neighbor (k-NN), and Decision Tree (DT).

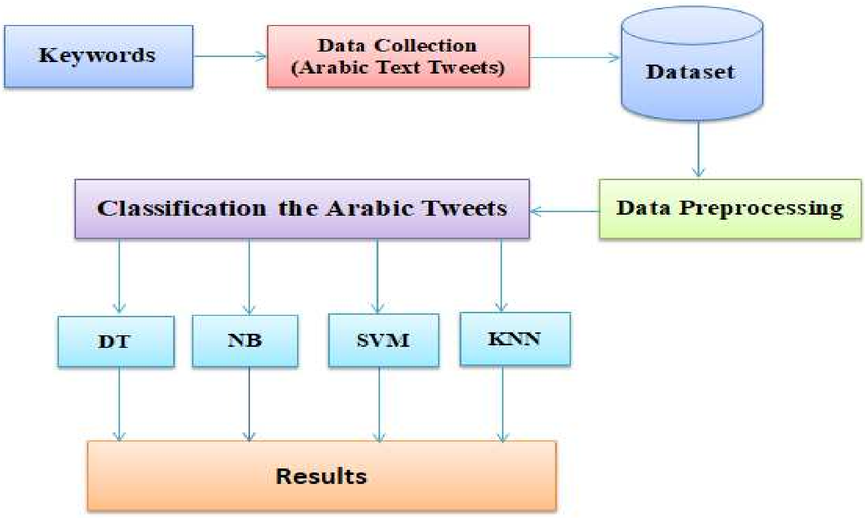

Figure 1: Flow Chart of the proposed system.

3.1 Data Collection

We collected the data from Twitter by collecting Arabic language tweets from several Arab countries and using the following steps: First, we created a Twitter account to be able to collect tweets using Twitter API. In this study, we have collected 54065 Arabic tweets along with their locations that are related to the influenza disease by using around 34 keywords such as الصداع which means

”headache” فلوزت which means” I’m suffering flu”, انا مرشح which means ”I have flu”. The dataset was annotated manually. Each tweet is either classified into ”valid” or ”invalid”. The valid category indicates the tweets that are related to influenza. But the invalid category is for tweets that are not related to the influenza disease and that is out of the topic. Table 2 shows some examples of these categories.

The used keywords is redefined manually by using the most Arabic words related to the flu symptoms found in several medical sites. Also, we utilized the Google trends site to refine the list and using the following steps: We find the top related search queries for each keyword. First, we enter the Arabic word related to flu disease in the search. Then, the Google trends site returns the most words related to the word given in specific Arab countries in a specific time of year.

The number of tweets is reduced from 54056 tweets to 6300 through two steps. The first step is done automatically using the Excel program by deleting duplicates tweets. The second step is done manually because Excel cannot delete the duplicate tweet if it has more space or letter or emotions. In the first step, the number of tweets is reduced to the 22000 tweets, while in the second, the number is reduced to 6300 (1473 for the valid category and 4827 for the invalid category).

<table>
<tr><td rowspan="5">Valid Category</td><td>No.</td><td>Arabic Example and Translation</td></tr>
<tr><td rowspan="2">EV1</td><td>اسوأ الاشياء اللي مريت فيها هي فقدان الشهيه بسبب الرشح</td></tr>
<tr><td>One of the worst things I have experienced is anorexia due to the colds</td></tr>
<tr><td rowspan="2">EV2</td><td>فش نوم وبداية انفلونزا</td></tr>
<tr><td>No sleep and beginning of flu</td></tr>
<tr><td rowspan="4">Invalid Category</td><td rowspan="2">EiV1</td><td>حمئ الله السعوديه واهلها</td></tr>
<tr><td>God protects Saudi Arabia and its people</td></tr>
<tr><td rowspan="2">EiV2</td><td>هذه فتنة تسبب سيلان اللعاب</td></tr>
<tr><td>This temptation causes salivation</td></tr>
</table>

Table 2: shows examples for the categories of the Data

As shown in Table 2, there are two examples of valid tweets (EV1 and EV2) and two examples of invalid tweets (EiV1 and EiV2). In the EV1, the tweet indicates that the person is suffering from influenza; the word (”الرشح ” means ”colds”); this example refers to a valid tweet. Similarly, in the EV2 example, the tweet indicates that the person is suffering from influenza; the word (”بداية انفلونزا ” which means ” beginning of flu”); this example refers to flu. For the invalid examples as shown n EiV1 and EiV2. EiV1 indicates that the person is not suffering from influenza; the word (”حمئ ” which means ”protect”) in this example is not same to ("حُمى " which means "Fever"), which means is not referred to flu; so it is considered as invalid tweet. The EiV2 indicates that the person is not suffering from influenza; because the word ("سيلان اللعاب " which means "salivation") and the tweet meaning is not related to flu; hence this example is considered as invalid.

3.2 Preprocessing

Preprocessing is a step that is used to make the data ready for knowledge extraction. There are several preprocessing stages applied in the data like tokenization, filtering, and stemming. These stages are explained in detail as below:

- **Filtering**: this process is applied to delete unnecessary words that may exist as (iterative, punctuation marks, unwanted, and stop words). The most used popular filtering process is stop-words removal. Prepositions, conjunctions are considered as stop words [Allahyari et al., 2017]. In our study, this process has several steps. In the beginning, the first step aims to gather the tweet where each tweet will be in one line instead of being on more than one line. The second step aims to separate the tweets and their location in separate text files to facilitate the filtering process. The separating process is based on a specific keyword. The third step aims to deletes English alphabets, punctuation marks, symbols, and emotions. Finally, we put the tweets and their location in the Excel file and remove exact duplicate tweets and remove the semi-duplicate tweets manually. Then, we labeled them as valid or invalid. The Labeling is based on the tweets whether related to influenza or not. Then, we used RapidMiner tool [Kotu and Deshpande, 2014] by depending on two parameters: filtering stop words and building n-Grams operator which will happen after the tokenization process.

- **Tokenization**: this process aims to divide sentences into chunks, whether that is words or phrases, and produces smaller pieces which are called tokens. It could be based on punctuation marks or whitespace [Allahyari et al., 2017]. There are different choices for dividing the sentences in RapidMiner, it can be presented with three options: mode, characters, and expression [Verma et al., 2014]. In this study, we depended on a mode-parameter. Nonletters were the default value which is used for splitting the Arabic tweets [Verma et al., 2014]. After that, the tokens are utilized for further processing stages [Pratama and Sarno, 2015], and [Attia, 2007].

- **Stop words removal**: This task is used to delete unnecessary and meaningless words such as (، to), (, on). Stop words return to words which are repeatedly used in an Arabic document [Ahmed et al., 2018], [Wahbeh et al., 2011]. Every token is compared with the existing stop word list. If it matches the list then it is deleted [Verma et al., 2014].

- **N-Grams**: it is a technique that is used to extract keywords from the Arabic tweets. It is based on (n-number of tokens) which is used to keep the sentence at its own meaning. If n=2, it is called diagrams, and when n=3 it will be called trigrams [Ahmed et al., 2018], [Jivani et al., 2011]. For n=3, a sequence of three consecutive words (tokens) is generated for each Arabic tweet in the dataset. The default value of "n" in RapidMiner is 2. Hence, the accuracy is expected to be increased in the classification step [Verma et al., 2014].

- **Stemming**: Once all of the filtering steps are applied to the data. The data gets ready for the stemming process. The stemming process is a technique that is assigned to get the root from the derived words to return it into its origin [Pratama and Sarno, 2015], [Ahmed et al., 2018]. The stemming process has three types of algorithms: Statistical, Truncating and Mixed algorithms. Root Based and Light stemmers are the most commonly used approaches in the Arabic language. Porter is an example for the stemmer, which is used in the English language, Khoja is used in Arabic at is one of the Root based stemmer [Wahbeh et al., 2011]. In our study, we used root based stemmer to remove the suffixes and prefixes for Arabic tweets [Verma et al., 2014].

3.3 Classification

The preprocessed data were divided into two types: training dataset and testing dataset. Firstly, a training dataset is executed after the dataset preparation. The dataset includes the Arabic tweets and every tweet related to a specific label (valid or invalid). Secondly, a testing dataset is accomplished by testing the classifier model that built previously based on the unseen dataset.

In this paper, we applied several classification techniques that are built into RapidMiner tool [Kotu and Deshpande, 2014] such as NB, k-NN (value of k =5 & distance measure = Mixed Euclidean Distance), SVM (kernel type: dot), and DT (type: Decision Tree, maximal depth: 10 & criterion: gain_ratio) algorithms. In the next paragraph, we shed light on these algorithms and clarify them in detail as follows:

- **Decision Tree Algorithm (DT)** It is essentially a hierarchical tree that uses attribute value conditions in order to split the data. In another meaning, it is recursively splitting the training data into minimal parts by depending on a group of tests that are shown at every tree branch. The node in the tree is considered as a feature training test, every branch is sloping from the node matches to the feature value. An instance is categorized starting from the parent node, checking the feature of this node and traveling down the tree branch to the value of the feature for the specific instance. In the text status, the decision tree nodes provisions are usually known in the phrase in the script. Different techniques are used in the decision tree to enhance classification precision. In this work, we applied a DT with maximal depth: 10 and where selection criterion on which attributes are selected for splitting is based on gain_ratio.

– **Naïve Bayes Algorithm (NB)** It is one of the most vastly utilized in the sentiment analysis model. This algorithm is an overseen mode that depends on the previous learning before beginning the task. Also, it is based on a probabilistic algorithm and it is designed based on the Bayesian probability method. Firstly, when starting the sentiment analysis, The probability of each word is defined. After that, the classifier was constructed to grouping the tweets depending on labeling. The equation of Naïve Bayes is shown below in Equation 1:

$$P(H|X) = P(X|H)P(H)/P(X) \tag{1}$$

Where, *P(H)* is The probability of hypothesis H. *P(X)* is the probability of the evidence. *P (X|H)* is the probability of the *X* on *H* is true. *P (H|X)* is the probability of the *H* on *X*. The NB assumes that *T* is the dataset training, *X* includes *(X1. . . .Xn)*, *n* explains the attributes of the row. *L*resembles the regarding labels. If *L* is a convenient label for the current dataset, the classifier can refer *X*, belongs to the label with the highest probability.

– **Support Vector Machine Algorithm (SVM)** The SVM is a type of supervised machine learning, which means the class label is known. It is widely used in classification and regression problems. The aim of SVM is to find the maximum margin between the hyperplane and the points that are on the hyperplane boundary which is called the Maximum Marginal hyperplane (MMH). The points are called Support Vector [Allahyari et al., 2017]. If the SVM is single and the dataset has two attributes, the equation for the separated Hyper-plane as per the following as in Equation 2:

$$w0 + w1x1 + w2x2 = 0 \tag{2}$$

Where *w1* is the weight vector for the first attribute, *w2* is the weight vector for the second attribute and *w0* is a bias [Han and Kamber, 2003]. The equations for the sides of margin as per the following as shown in Equations 3, and 4:

$$H1 : w0 + w1x1 + w2x2 >= 1, yi = +1 \tag{3}$$

$$H2 : w0 + w1x1 + w2x2 <= -1, yi = -1 \tag{4}$$

Where yi is the class label for each tuple is a dataset that the values of it are 1 or -1 [Han and Kamber, 2003].

– **K-Nearest Neighbor Algorithm (*k*-NN)** It is a ranking method that utilizes a function that depends on the number of closest neighbors and the distance between the training data used to test the classification outcomes. The cosine similarity is the distance function that has been utilized to complete the test; it is vastly applied in the datasets to detect the similarity among different texts. The class in the text is detected by electing on K nearest neighbor and this neighbor is defining the highest value. In this work, the value of k is set to be 5 and the applied distance measure is Mixed Euclidean Distance.

4 The Experimental Results and Discussion

In this work, firstly, we applied the preprocessing techniques on the collected data as shown in Section 3.2. Then, several classifiers were applied which are: SVM, NB, DT, and k-NN. We validated the models by using the cross-validation technique that divides the dataset into a training dataset and testing dataset based on the k-fold value; to avoid the dataset overfitting and to enhance the model performance. Different folds values are applied which are 5, 10, 15 and 20. There are many metrics were used to evaluate the results and compared between the classifiers [Han and Kamber, 2003]. These metrics are accuracy, precision, recall and F1-measure; these metrics are calculated based on the following Equations 5,6,7, and 8:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{5}$$

$$Precision = TP/(TP + FP) \tag{6}$$

$$Recall = TP/(TP + FN) \tag{7}$$

$$F1 - measure = 2 * ((Precision * Recall)/(Precision + Recall)) \tag{8}$$

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).

In this work, we have conducted three experiments on the data. The first experiment was applied to the tweets which are related to the Arab countries and using multiple folds values. The second experiment is applied to the tweets based on the country name. The third experiment is applied to the tweets which are related to the countries with the same Arabic accents. In the next section, the experiments are clarified in detail.

4.1 First Experiment

The first experiment used multiple folds that were selected as (5, 10, 15 and 20). Then, we applied them to all the labeled tweets without using GIS. Each time the tweets are tested with different fold and different classifiers such as k-NN, NB, DT, and SVM. Figure 2 displays the average accuracy for different folds which are: (5, 10, 15, and 20) using the different classifiers. The results are as follows: the average accuracy of the k-NN 82.61%, 82.61%, 82.72%, and 82.88% respectively. The average accuracy of the NB is 82.01%, 82.64%, 82.77%, and 83.20% respectively. Also, the average accuracy of the DT is 82.84%, 82.45%, 82.88%, and 82.48% respectively. Finally, the average accuracy of the SVM is 81.45%, 82.21%, 82.21%, and 82.61% respectively. After comparing all the accuracy values, we noted that the best accuracy value was 83.20%, which is achieved using the NB algorithm at 20-folds. We also noted that the accuracy value is approximately constant at a specific value; because when the k-fold value increases the accuracy is nearly getting stable and this agreed with the results in [Moss et al., 2018].

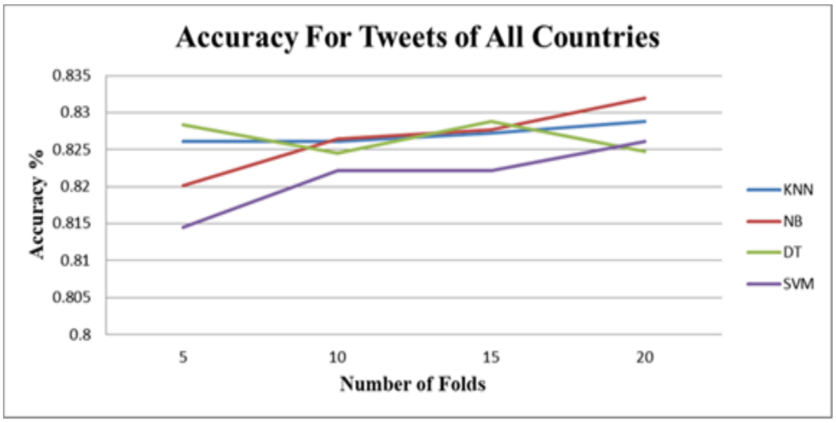

Figure 2: Accuracy for Tweets of All Arabic Countries based on the number of folds and for the four classifiers(k-NN, NB, DT, and SVM).

4.2 Second Experiment

The second experiment has been applied to all the classifiers at a fixed k-fold value which is 20 (chosen based on the previous experiment as in 4.1). This experiment is applied based on the Arab countries, which we collected the data from, namely Algeria, Bahrain, Egypt, Iraq, Jordan, KSA, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Sudan, Syria, Tunisia, UAE, and Yemen. All of the previous countries were applied except Lebanon, Morocco, and Mauritania due to less shared Arabic tweets related to influenza 23, 11 and 7 tweets respectively; we excluded these countries for this experiment. Table 3 shows the accuracy values for the Arab countries using all classifiers at 20-folds. We noted that the best accuracy value was achieved for Syria country, which is 89.06% at the NB classifier.

Country	Classifiers			
Name	k-NN	NB	DT	SVM
Algeria	80.98	80.13	80.47	75.07
Bahrain	86.14	84.07	**88.74**	76.34
Egypt	84.4	83.04	81.77	80.09
Iraq	84.22	79.86	84.18	77.1
Jordan	84.04	81.44	81.03	74.12
KSA	**87.9**	83.31	81.67	79.83
Kuwait	87.26	85.74	88.37	76.74
Libya	78.46	85.96	72.37	69.4
Oman	80.62	79.23	73.77	72.84
Palestine	82.71	79.00	75.44	68.42
Qatar	85.46	83.64	85.44	82.43
Sudan	86.73	87.23	88.64	78.86
Syria	84.22	**89.06**	87.00	**85.89**
Tunisia	85.48	85.54	79.79	72.83
UAE	86.31	86.3	80.31	75.02
Yemen	80.79	81.84	77.11	73.68

Table 3: The Accuracy for the Arab countries using all classifiers

Table 4 shows the Precision values for the Arab countries using all classifiers at 20-folds. We noted that the best precision value was in Syria, which is 93.405% at the DT classifier. Table 5 shows the Recall values for the Arab countries using all classifiers at 20-folds. We noted that the best Recall value was in Tunisia, which is 82.72% at the NB classifier. Table 6 shows the F1-Measure values for the Arab countries using all classifiers at 20-folds. We noted that the best F1-Measure value was in Bahrain, which is 83.398% at the DT classifier.

While there are different values for the accuracy in different classifiers, we noted that the SVM classifier always has the worst accuracy value. Syria was having the highest accuracy value at the NB classifier which is 89.06%.

Country	Classifiers			
Name	k-NN	NB	DT	SVM
Algeria	74.635	73.525	78.95	37.5
Bahrain	81.355	78.025	88.68	38.145
Egypt	78.565	74.005	82.525	**90**
Iraq	82.375	71.9	86.415	38.54
Jordan	79.41	76.25	77.36	81.475
KSA	84.245	75.435	83.51	85.79
Kuwait	**85.565**	79.96	89.605	88.335
Libya	74.98	**84.445**	70.535	48.21
Oman	75.865	74.58	66.73	86.345
Palestine	81.045	76.25	79.115	34.18
Qatar	76.115	71.16	79.955	41.33
Sudan	83.385	71.46	91.715	89.2
Syria	60.095	77.385	**93.405**	42.935
Tunisia	82.325	81.405	85.465	36.39
UAE	85.22	82.33	78.45	87.44
Yemen	75.48	76.565	73.27	36.84

Table 4: The Precision for the Arab countries in all classifiers

Country	Classifiers			
Name	k-NN	NB	DT	SVM
Algeria	73.675	70.83	63.825	50
Bahrain	78.895	77.165	**78.71**	50
Egypt	67.64	76.135	56.145	51.035
Iraq	69.15	73.9	67.03	50
Jordan	78.99	79.46	65.835	50.6
KSA	77.02	80.32	57.55	52.345
Kuwait	78.015	81.715	77.125	50.59
Libya	72.43	81.565	57	50
Oman	75.04	77.43	56.36	50.835
Palestine	77.805	78.715	62.6	50
Qatar	65.805	69.625	61.565	49.815
Sudan	75	80.505	74.7	**52.805**
Syria	53.87	82.425	53.845	50
Tunisia	**79.805**	**82.72**	63.52	50
UAE	76.815	80.775	64.735	50.945
Yemen	77.2	77.715	60.035	50

Table 5: The Recall for the Arab countries in all classifiers

4.3 Third Experiment

In the third experiment, we applied the classifiers at a fixed k-fold value which is 20 and using the Arabic region and based on the accent as per the geographic location with the GIS. We divided the Arab countries into five regions based on the assumption which are: Arab Maghreb States, Iraq, Levant, Nile Basin countries, and the Arabian Gulf. Figure 3 shows that the Arab World Regions accuracy values for all classifiers at 20-folds. We noted that the best accuracy value was in the Arabian Gulf, which is 86.43% using k-NN classifier.

Figure 4 shows the precision values for the Arab World Regions using all classifiers at 20-folds. We noted that the best value was in the Nile Basin Gulf,

Country	Classifiers			
Name	k-NN	NB	DT	SVM
Algeria	74.15	72.15	70.59	42.86
Bahrain	80.11	77.59	**83.40**	43.28
Egypt	72.69	75.05	66.83	65.13
Iraq	75.19	72.89	75.50	43.53
Jordan	79.20	77.82	71.13	62.43
KSA	80.47	77.80	68.14	65.02
Kuwait	**81.62**	80.83	82.90	64.33
Libya	73.68	**82.98**	63.05	49.09
Oman	75.45	75.98	61.11	63.99
Palestine	79.39	77.46	69.90	40.60
Qatar	70.59	70.38	69.57	45.18
Sudan	78.97	75.71	82.34	**66.34**
Syria	56.81	79.83	68.31	46.20
Tunisia	81.05	82.06	72.88	42.12
UAE	80.80	81.55	70.94	64.38
Yemen	76.33	77.14	66.00	42.42

Table 6: The F1-Measure for the Arab countries in all classifiers

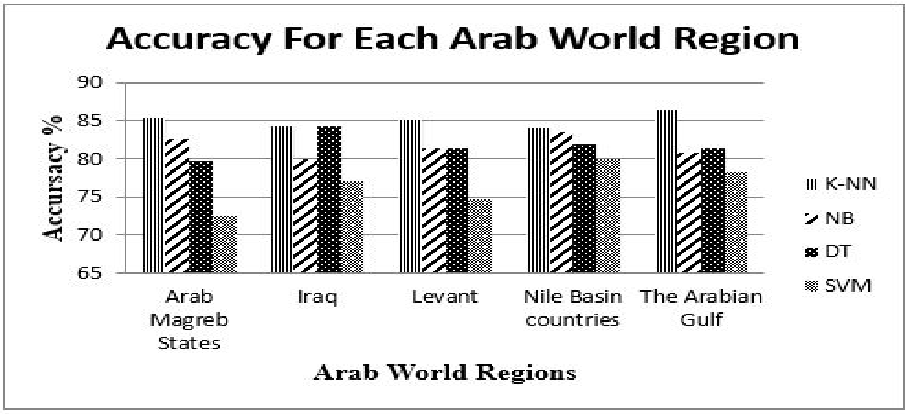

Figure 3: The Accuracy for the Arab World Regions in all classifiers.

which is 89.905% at the SVM classifier.

Figure 5 shows the Recall values for the Arab World Regions using all classifiers at 20-folds. We noted that the best value was in the Arab Maghreb States, which is 80.93% at k-NN classifier.

Figure 6 shows the Arab World Regions F1-Measure values for all classifiers at 20-folds. We noted that the best value was in the Arab Maghreb States, which is 81.61% at k-NN classifier.

While there are different values for the accuracy, precision, Recall and F1-Measure using different classifiers, we noted that the SVM mostly has the worst values. The Arabian Gulf has the highest accuracy value at k-NN classifier, which was 86.43%.

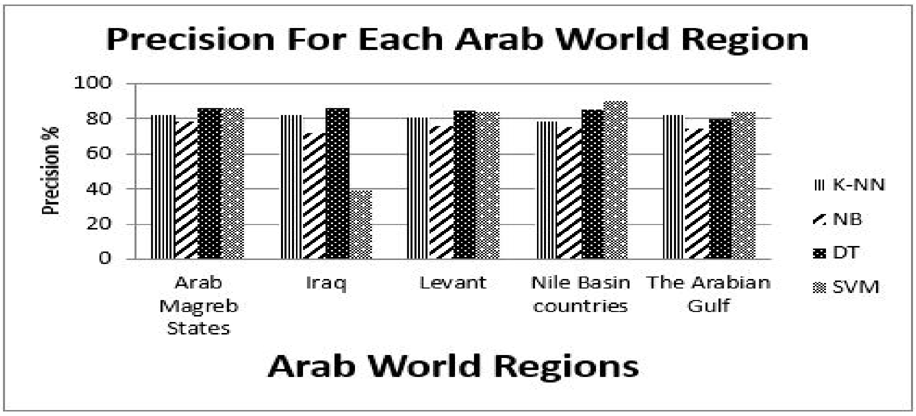

Figure 4: The Precision for the Arab World Regions in all classifiers

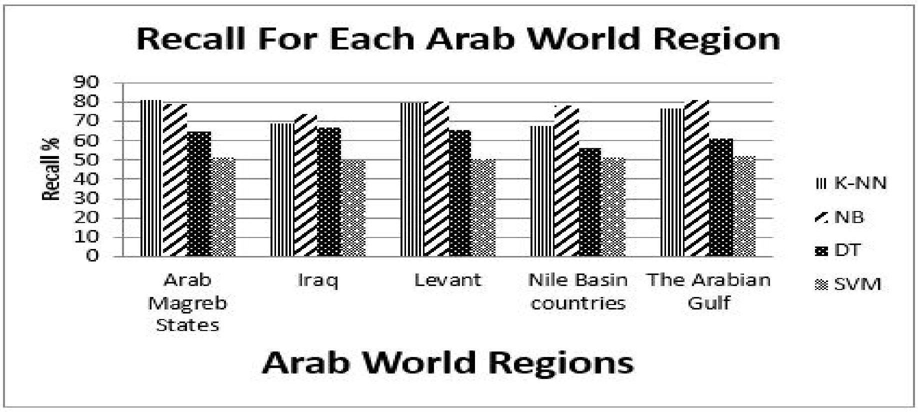

Figure 5: The Recall for the Arab World Regions in all classifiers

We summarize the experimental findings and results in the following:

- The best classifier that gives the highest average accuracy for different k-fold (5, 10, 15, and 20) for all tweets of all Arab countries is the NB with 83.20% at 20-folds based on the first experiment as shown in Figure 2.
- The best classifier that gives the highest average accuracy for Arab country's tweets individually at 20-folds is NB with 89.06% accuracy of Syria based on the second experiment as shown in Table 3. This means that Syria is the most active country used Twitter to detect influenza disease epidemics.
- The best classifier that gives the highest average accuracy for Arab country's tweets individually at 20-folds is NB with 86.43 % accuracy of the Arabian

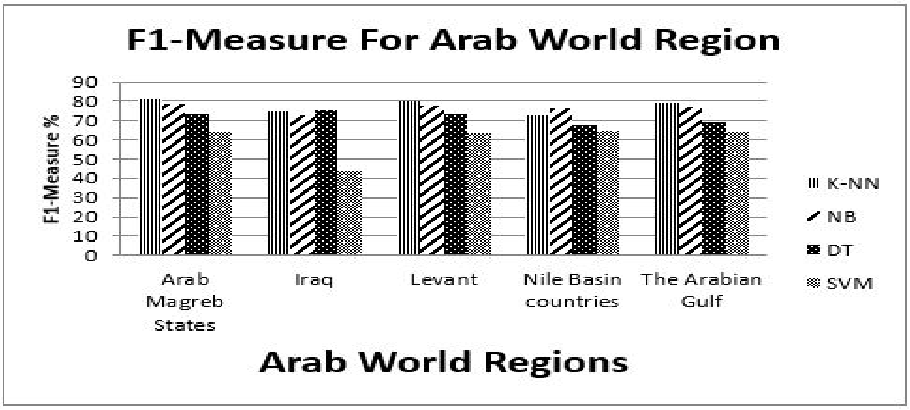

Figure 6: The F1-Measure for the Arab World Regions in all classifiers

Gulf based on the third experiment as shown in Figure 3. This means that "The Arabian Gulf" countries are the most active region which has used their own accent to type the Arabic text on twitter to detect the disease epidemics.

– The DT algorithm is suitably used in the small dataset. For this reason, the DT gives better performance than the k-NN. Because the number of tweets in each Arab country is less than the number of tweets in each Arab region as shown in Table 3. But the k-NN and NB algorithms are suitably used in large datasets. For this reason, the k-NN and NB give better performance with DT. Because the number of tweets in each Arab region is greater than the number of tweets in each Arab country as shown in Figure 3.

5 Conclusions

This paper provides an approach to detect Influenza disease epidemics by classifying the Arab community tweets in the Arabic language. A vast amount of tweets was collected and preprocessed. The collected tweets were manually labeled into two labels: valid or invalid and each tweet is matched with the proper label. Several classifiers (NB, k-NN, SVM, and DT) were applied to the data. Different k-fold (5, 10, 15 and 20) were used to determine the best k experimentally. The Accuracy values from each experiment were calculated to evaluate the performance of the proposed system for each classifier at each fold. We found that the best accuracy value was 83.20%, which was achieved in the NB algorithm at 20-folds. The 20-folds value is selected because the accuracy is the best and it approximately getting stable after this value. Syria had the best accuracy, amongst other countries, which was 89.06% at the NB algorithm. The Arabian

Gulf had the highest accuracy value, amongst other regions, at the k-NN algorithm which was 86.43%. We found that the SVM classifier produced the worst results in all of the experiments. These results proved that the Arabic tweets have shown the emergence and occurrence of influenza amongst people. This, in particular, indicates that the Arab countries can actively help in fighting the influence of influenza before they possibly occur in the neighboring countries or districts. It also proved that data mining techniques can successfully extract useful information and generate results that can have a substantial impact on the performance and decision making in the future.

6 Future Work

In the upcoming future, we will try to increase the prediction accuracy of the collected data by using deep learning techniques and other natural language processing approaches. Also, we can gather local or regional reports which are related to the influenza outbreak and compare it to the Twitter text mining results. Also, we may collect data using online questionnaires and ask people to fulfill it; to delve deeper into the situation and seasons in which the outbreak may take place.

Acknowledgment

Thanks to Jordan University of Science and Technology for supporting this publication under Award Number 20170030.

References

[Ahmed et al., 2018] Ahmed, W., Bath, P. A., Sbaffi, L., and Demartini, G. (2018). Moral panic through the lens of twitter: An analysis of infectious disease outbreaks. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 217–221. ACM.

[Al-Zinati et al., 2019] Al-Zinati, M., Almasri, T., Alsmirat, M., and Jararweh, Y. (2019). Enabling multiple health security threats detection using mobile edge computing. *Simulation Modelling Practice and Theory*, page 101957.

[Alessa and Faezipour, 2018] Alessa, A. and Faezipour, M. (2018). A review of influenza detection and prediction through social networking sites. *Theoretical Biology and Medical Modelling*, 15(1):2.

[Allahyari et al., 2017] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

[Allen et al., 2016] Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., and Gawron, J.-M. (2016). Applying gis and machine learning methods to twitter data for multiscale surveillance of influenza. *PloS one*, 11(7):e0157734.

[Aramaki et al., 2011] Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics.

[Aslam et al., 2014] Aslam, A. A., Tsou, M.-H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., Peddecord, K. M., Nagel, A. C., Allen, C., Yang, J.-A., et al. (2014). The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of medical Internet research*, 16(11):e250.

[Attia, 2007] Attia, M. A. (2007). Arabic tokenization system. In *Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources*, pages 65–72. Association for Computational Linguistics.

[Bernard et al., 2018] Bernard, R., Bowsher, G., Milner, C., Boyle, P., Patel, P., and Sullivan, R. (2018). Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks. *Journal of Public Health*, 26(5):509–514.

[Chew and Eysenbach, 2010] Chew, C. and Eysenbach, G. (2010). Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118.

[Culotta, 2010] Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. acm.

[Culotta, 2013] Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Language resources and evaluation*, 47(1):217–238.

[Fung et al., 2013] Fung, I. C.-H., Fu, K.-W., Ying, Y., Schaible, B., Hao, Y., Chan, C.-H., and Tse, Z. T.-H. (2013). Chinese social media reaction to the mers-cov and avian influenza a (h7n9) outbreaks. *Infectious diseases of poverty*, 2(1):31.

[Han and Kamber, 2003] Han, J. and Kamber, M. (2003). Classification and prediction, data mining: Concepts and techniques.

[Jivani et al., 2011] Jivani, A. G. et al. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6):1930–1938.

[Kim et al., 2013] Kim, E.-K., Seok, J. H., Oh, J. S., Lee, H. W., and Kim, K. H. (2013). Use of hangeul twitter to track and predict human influenza infection. *PloS one*, 8(7):e69305.

[Kotu and Deshpande, 2014] Kotu, V. and Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

[Lee et al., 2013] Lee, K., Agrawal, A., and Choudhary, A. (2013). Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1474–1477. ACM.

[Lee et al., 2017] Lee, K., Agrawal, A., and Choudhary, A. (2017). Forecasting influenza levels using real-time social media streams. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 409–414. IEEE.

[Moss et al., 2018] Moss, H. B., Leslie, D. S., and Rayson, P. (2018). Using jk fold cross validation to reduce variance when tuning nlp models. *arXiv preprint arXiv:1806.07139*.

[Pratama and Sarno, 2015] Pratama, B. Y. and Sarno, R. (2015). Personality classification based on twitter text using naive bayes, knn and svm. In *2015 International Conference on Data and Software Engineering (ICoDSE)*, pages 170–174. IEEE.

[Quwaider and Jararweh, 2016] Quwaider, M. and Jararweh, Y. (2016). A cloud supported model for efficient community health awareness. *Pervasive and Mobile Computing*, 28:35–50.

[Santos and Matos, 2014] Santos, J. C. and Matos, S. (2014). Analysing twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling*, 11(1):S6.

[Signorini et al., 2011] Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

[Smadi and Qawasmeh, 2018] Smadi, M. and Qawasmeh, O. (2018). A supervised machine learning approach for events extraction out of arabic tweets. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 114–119. IEEE.

[St Louis and Zorlu, 2012] St Louis, C. and Zorlu, G. (2012). Can twitter predict disease outbreaks? *Bmj*, 344:e2353.

[Suarez et al., 2018] Suarez, D., Araque, O., and Iglesias, C. A. (2018). How well do spaniards sleep? analysis of sleep disorders based on twitter mining. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 11–18. IEEE.

[van de Belt et al., 2018] van de Belt, T. H., van Stockum, P. T., Engelen, L. J., Lancee, J., Schrijver, R., Rodríguez-Baño, J., Tacconelli, E., Saris, K., van Gelder, M. M., and Voss, A. (2018). Social media posts and online search behaviour as early-warning system for mrsa outbreaks. *Antimicrobial Resistance & Infection Control*, 7(1):69.

[Verma et al., 2014] Verma, T., Renu, R., and Gaur, D. (2014). Tokenization and filtering process in rapidminer. *International Journal of Applied Information Systems*, 7(2):16–18.

[Wahbeh et al., 2011] Wahbeh, A., Al-Kabi, M., Al-Radaideh, Q., Al-Shawakfa, E., and Alsmadi, I. (2011). The effect of stemming on arabic text classification: an empirical study. *International Journal of Information Retrieval Research (IJIRR)*, 1(3):54–70.

[Wang et al., 2018] Wang, J., Zhao, L., Ye, Y., and Zhang, Y. (2018). Adverse event detection by integrating twitter data and vaers. *Journal of biomedical semantics*, 9(1):19.

[Ye et al., 2016] Ye, X., Li, S., Yang, X., and Qin, C. (2016). Use of social media for the detection and analysis of infectious diseases in china. *ISPRS International Journal of Geo-Information*, 5(9):156.