# Tracking Coronavirus Pandemic Diseases using Social Media: A Machine Learning Approach

Nuha Noha Fakhry[1], Gamal Kassam[3]
Business Informatics
German University in Cairo, Cairo, Egypt

Evan Asfoura[2]
Marketing Department, College of Business
Dar Al Uloom University, Riyadh, Saudi Arabia

***Abstract*—With the increasing use of social media, a growing need exists for systems that can extract useful information from huge amounts of data. While, People post personal data interactively, an outbreak of an epidemic event can be noticed from these data. The issue of detecting the route of pandemic diseases is addressed. The main objective of this research work is to use a dual machine learning approach to evaluate current and future data of Covid-19 cases based on published social media information in specific geographical region and show how the disease spreads geographically over the time. The dual machine learning approach used based on traditional data mining methods to estimate disease cases found in social media related to specific geographical region. On other hand, sentiment analysis is conducted to assess the public perception of the disease awareness on the same region.**

***Keywords*—Pandemic diseases; outbreak detection; social media; sentiment analysis; machine learning; text mining; geo-located data; CRISP-DM**

## I. INTRODUCTION

Nowadays, a lot of factors are responsible for the emergence of infectious diseases such as SARS or avian influenza, which leads to more health challenges worldwide, to avoid these concerns government and global health leaders worked together and revised the IHR. The IHR provides the framework and regulations to detect and respond to outbreak diseases [1]. In 2007, All countries that had knowledge about an outbreak of a disease with an international concern had to report it to the WHO within 24 hours of notice to help in the containment of the disease and to be timely efficient regardless of the location that the disease spread in. However, this information may not be sufficient enough to respond to any disease outbreak [2].

New technologies lately that are internet related, change the way people access or find out about health information. Official health authorities do not fully control the information related to a certain new epidemic disease as journalists and the general public are now in direct contact with raw data. Social media allows reporting mechanisms that makes it easier for the public to access it [1]. Social media such as Twitter and Facebook are increasingly being used as tools for real-time knowledge discovery relating to social events, emerging threats, epidemics, and even product trends [3]. For example, real time analysis of Twitter users' tweet content can be or is being used to detect earthquakes and provide warnings [4], to identify needs (e.g., medical emergencies, food and water shortages) during recovery from natural disasters such as the Haiti Earthquake [5], and to track emergence of specific syndromic characteristics of influenza-like illness [6], and collect epidemic related tweets [4]. The role of social media in the biomedical domain has become significant in recent years [3, 7]. Researchers and physicians have utilized social media data to communicate and share information between patients and health care decision makers [3], and to develop large scale, dynamic disease surveillance systems [7] and mining biomedical and health-related information [4]. An immediate and direct use of social media in the biomedical domain is a mean for patients and professionals to communicate and exchange information. Web 2.0 along with ubiquitous mobile computing devices allows individuals to dynamically and seamlessly interact with each other in real time, regardless of their locations. PatientsLikeMe is a social network for patients that improves lives and a real-time research platform that advances medicine. On PatientsLikeMe's network, patients connect with others who have the same disease or condition, allowing them to track and share their own experiences. Eijk et al. illustrated the use of OHCs for ParkinsonNet, a social network for Parkinson disease patients whose participants (both patients and professionals) use various types of OHCs to deliver patient-centered care [8]. Merolli et al. explored different ways that chronic disease sufferers engage in social media in order to better tailor these online interventions to individually support patients in specific groups [9]. Additionally, Twitter, Facebook, and other social blogging services provide conduits for patients and medical practitioners to collaborate, exchange, and disseminate information through official broadcasting channels/webpages or discussion groups [10]. Most popular social media providers such as Twitter and Facebook allow their posts to be geo-located. These properties provide researchers in the healthcare community the ability to monitor the medical related emergencies. In late 2019, a novel coronavirus was identified (COVID-19). This is likely the third time in three decades that a zoonotic coronavirus has jumped from infecting animals to humans. As of today, 213 have died and 9809 have been infected in China with the center of the outbreak being located in Wuhan but now having spread, with confirmed cases in twenty-two other countries. While the fatality rate of COVID-19 is less than other recent respiratory virus outbreaks, it remains much higher than other commonly encountered causes of respiratory infection but its full impact is yet undetermined. The WHO has now declared the coronavirus outbreak to be a public-health emergency of international concern [11].

The aim of this work is to be able to track pandemic diseases using machine learning techniques, text mining and sentiment analysis through social media platforms. The solution will follow the CRISP-DM process which provides a structured approach to planning a data science project, it'll be explained in detail in the methodology section. This paper is organized as follows, the first section discusses machine learning and its techniques. The second section includes an introduction to sentiment analysis and how it could be achieved. The third and last section discusses the role that social media plays in public health.

## II. BACKROUND

There are many Challenges and opportunities of Social media in the Public health is known to be 'the science and art of preventing disease, prolonging life and promoting health through the organized efforts and informed choices of society, organizations, public and private, communities and individuals. Information is the main source to the public health while health data is the foundation. The timelines of health data controls and limits the actionable information of the public health as the conventional route of the data goes all the way from the patient's self-report to the doctor, diagnosis is confirmed and then data goes from the doctor or the facility to the public health authority. Health data that are present on social media differ from the mentioned conventional route by excluding the 'middleman'. Unfortunately, the middle man has an important role. They confirm information about the population and send them to the authorities they serve in order to increase the level of confidence of the information. Governmental actions or interventions are dependent on this confidence. Back in 2009, in the outbreak of H1N1 and Haiti cholera the public health officials realized that social media has the ability to indicate the trends of disease outbreak faster and with higher quality in comparison with traditional methods of public health reporting. Mining online data, searching the behavior of the internet data and social network data began from researchers and practitioners. The purpose was to help in predicting a variety of social, economic, behavioral and health-related events. Majority of the work was on predicting aggregate properties, like the commonness of occasional influenza in a given area, country or a city (e.g. Google Flu Trends, Monitoring Dengue activity using Internet search [12, 13].

There are also challenges that face social media in disease surveillance. The most asked question is how to make sure of the verification of the information coming from social media, as the verification of such large noisy data is considered a big challenge. Public health officials integrate information coming from social media as it could add additional advantages to their surveillance responsibilities [14]. However, analyzing social media data can be done through verifying and comparing them with other sources, which will help in identifying rumors early. A challenge concerning this issue, the spread of rumors across multiple social media platforms or a malicious actor spamming the system with false information, this makes validation more difficult. To overcome these issues many systems verify the messages by reviewing them through a moderator, reports to be labeled clearly as community contributions and enable user's feedback and corroboration of submissions, this was proven to be successful by Wikipedia. In general, social media is considered to be a two-way information exchange as the crowd can see and evaluate the quality of information published by other users. Many challenges exist in mining the social media. First, data that consists of text might be hard to analyze as harvested data (e.g. a tweet) may not contain complete information and meaning that helps in the classification process. Also, coding for geographic origins may have certain limitations, accounts on social networking sites may not contain geographic information so visible geographic information won't be accessible [15]. So it's better to use data mining sources that track IP addresses or use techniques to monitor social media activity on mobile phones. The majority of new mobile phones have Global Positioning Systems (GPS) or monitoring chips that can easily be attached to independent devices. This technology can track the location of the device needed to be tracked, but also as a challenge users may now allow to share this information for public use. As social media platforms are in constant growth, the attention should be focused on the potential demographic biases coming from users of any service. Moreover, overfitting is a famous issue in machine learning, it's a used methodology in mining large data sets like those extracted from social media. Constructing models that fit many data points coming from social media is easy in order to calculate statistics from public health sources (e.g. disease incidence curves), predictability of these models should be tested. This issue can also be addressed by avoiding using data coming from official sources in the process of developing the models. For example, in article [16], sentiments about vaccinations were extracted by performing fraction classification on the data manually, then machine learning algorithms are applied on the trained on human-labeled data in order to evaluate all of the remaining data in the data set. After developing the fully labeled data set, correlation was done against the estimates of the CDC regarding H1N1 vaccination rates per geographic region, estimating the strong correlation between sentiments and vaccination rates (in a certain direction, e.g. regions with more positive sentiment had higher vaccination coverage).

This work aims to cover the lake of the previous researches in machine learning field using Twitter as a social media platform in order to respond to the pandemic crisis happening right now worldwide, no articles were found to have used machine learning techniques in applying it for detecting people that have symptoms related to the novel Coronavirus. Social media platforms are considered to be a huge source of data especially that it provides real-time data. There was a gap, no articles used Social media platforms in order to detect illness and make future predictions in order to support hospitals nowadays. Article [8] presented an automatic detection model for Coronavirus from X-ray images utilizing transfer learning with convolutional neural networks. Another article [17] used machine learning methods in order to classify Coronavirus using CT Images, this is more of a diagnostic solution.

In order to solve the problem and also accomplish the mentioned opportunities in the introductory chapter in section 1.3, a model was built. Extracting data from Twitter which is a

popular social media platform and classifying them using machine learning techniques. A CRISP DM approach was used.

## III. Research Methodology

The model used in this Paper follows the same framework in Fig. 1. The model begins with the data extraction phase from any social media platform then the data extracted follows two paths.

The first path includes a couple of data preprocessing and tagging techniques then a classification algorithm should be applied to the preprocessed data. The results should be visualized in order to observe patterns that will lead to the outbreak prediction or tracking. The second path is related to the process of sentiment analysis. The data extracted should undergo the same preprocessing steps, perform sentiment analysis on the extracted data using a machine learning approach and finally assess the statistics and visualize them

This model help to raise awareness about how the Coronavirus is spreading in a statistical manner all over the United States. This will be achieved by extracting twitter data about the virus and trying to detect from the text if someone is sick or tested positive for the virus. The classified data will help hospitals to expect a certain amount of cases per day and be prepared for the future predictions made. Also, sentiment analysis using machine learning was made in order to analyze people's consent about the pandemic. A Quantitative method is used as it will maximize the results of our findings, as well as facilitating predictions which is one of the important goals of this paper. Also, quantitative research involves assigning numbers to variables which are used later in extracting statistical information related to these variables as well as exploring the relationship between them that can be used to compare the variables to real time data.

The work's solution follows a CRISP-DM approach as mentioned in the introduction. The process life cycle follows six phases. The first phase is business understanding, this phase is related to the business point of view and understanding the target of the work and how to get there through designing a plan to accomplish the objectives. The second phase is data understanding, this phase is concerned with collecting the data, getting familiar with it and discovering interesting patterns. The third phase is data preparation, this phase includes processes that are applied on the data to form the final dataset that enters the model for example, data cleaning, generating new attributes and so on. The fourth phase is the modeling, in this phase a couple of approaches are applied as some problems in data mining have more than one solution approach. The fifth phase is evaluation, in this stage one of the approaches implemented in the modeling phase wins as it ensures high quality but before going to the final phase this model should be evaluated to be sure it achieves the work's objectives. The sixth and final phase is deployment, this phase doesn't mean that an end for the work is reached as the information gained has to be visualized in a way that whoever is going to use it will be able to understand it. It's important to decide on what actions will be done in order to make use of the implemented model. [18].

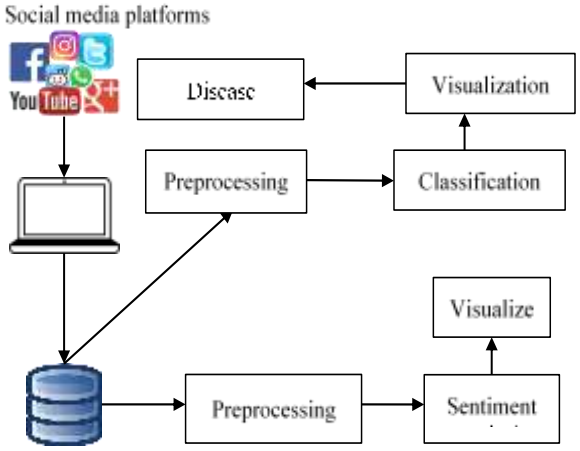


Fig. 1. Conceptual Design.

Some external software components will be used in this work which are of great use in the machine learning, text mining and sentiment analysis fields. Twitter API account is used to give access in gathering tweets from Twitter. As well as R, which is a free software under the GNU-license that allows us to perform a variety of different computations and visualizations on data. In a way it's quite similar to MATLAB. R contains a lot of different mathematical functions for dealing with time series, such as the possibility to automatically compute the autocorrelation function of a time series and visualize it in a compelling manner. All plots of time series data in this work have been created using R, Lastly RapidMiner which is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. Extensions were added to RapidMiner to be used in this work: MeaningCloud text analytics, Text analysis by AYLIEN, Rosette text analytics, text processing and finally WordNet extension. Using these tools will help us gather the data that'll support this work as well as understanding the data and exploring the hidden patterns to be able to provide useful information that people, hospitals and States will benefit from.

## IV. Implementation

This work follows the CRISP-DM phases which are discussed in the methodology section. This section will provide the detailed procedures followed in the work that helped in reaching the stated goals and solving the problems mentioned in the introduction.

### A. Business Understanding

This part includes the objectives and goals of the work to be able to identify the problems related to data mining as well as the opportunities that this work should handle. The problems are how to extract tweets from Twitter and be able to classify them based on whether a person is sick or not and also after the classification of the tweets they should be

compared to real world data in order to compare the results and measure their accuracy. If these two problems are solved, a couple of opportunities will arise, the first one is if the data is correlated with the real-world data this will be faster in tracking the disease and be able to help people, hospitals or countries allocate the needed resources for the current crisis. Also, future predictions can be made from patterns extracted from the data and lastly, the data gathered can be used to measure the sentiment of people and see how they're reacting towards the crisis and whether they need mental health support or not.

### B. Data Understanding

The data used in this work were tweets posted during the Coronavirus crisis event. Twitter API makes it simple to gather data from twitter using keywords, language, geolocation and more. In order to gather data from Twitter API, a simple code was written in R to support the gathering process. Several data sets are gathered during the month of June in the year 2020. After the data is gathered it's being stored in an excel sheet which is going to act as the database for simplicity. As soon as the data is stored the preprocessing and tagging begins and also sentiment analysis will be performed to classify positive and negative tweets. The main objective is to perform naive bayes analysis to be able to predict from the tweet if someone is actually sick or tested positive for the Coronavirus, also the keywords that influenced the classification and finally comparing the accuracy of the classifiers together. A lot of tweets are somehow dependent on the history for example if someone tweeted "I'm very sick right now" this would be classified as a sick person tweet even if the user's history indicated that he suffers from cancer as i won't be considering the history of tweets as it requires a more sophisticated system.

In the initial approach section random tweets were extracted from all around the world from R related to Coronavirus using specific keywords. The first search query contained the keyword "Corona". The second search query contained "cough OR fever OR tiredness". Finally the third search query contained "corona AND tested AND positive". Using data from all around the world was too general so instead data was extracted only from the 51 states in America. This was achieved by using a library in R that supports twitter API which is called (twitteR), this library had a function that helped in extracting tweets using geocode. It took four arguments which were latitude, longitude, a number that represents the area of a circle and the fourth argument was either Km or miles which depends on the area of the circle. Twitter then responds with data that has been geotagged only in this location. Unfortunately, when searching in a specific circle of radius it might as well include tweets from another state around the area so this tweet may appear twice in the data while searching in two different circles, to avoid this each tweet's individual id is checked and duplicates are removed.

This work focuses mainly on the spread of Coronavirus in the United States of America. A classifier cannot be used by multiple languages so in order to achieve this, tweets in English language were only extracted. The (twitteR) library in R has a function that supports extracting tweets in a specific language only. The positions of the extracted data from the United States are present in Table I.

### C. Rate Limiting of the Extracted Twitter Data

If the premium account is not purchased in Twitter API, unlimited requests cannot be made so a limit does exist. The limit is restricted to 180 GET requests per 15 minutes per user.

### D. Description of the Tagging Process

The twitter search API was used in order to extract the training dataset, tweets in English were extracted as many as possible by the process mentioned earlier in the description phase. Data was collected by searching for tweets using the keywords in section regional concerns as well as the specified tagged geo locations that were mentioned in the last section. The data was then stored in an xlsx format to be supported by RapidMiner.

Once the tweets were stored the tagging process started. Tweets that indicated that someone has tested positive for the virus or had symptoms were tagged as relevant, meaning that the label column was set to True while tweets that didn't indicate any of the cases were labeled False. It's important to put into consideration that all tweets that indicated the sickness of someone is classified as True, in other words if a mother tweeted that her son is sick or wishing someone to feel better, it will be classified as True.

In this work, only illnesses that had symptoms like fever, diarrhea, fever, tiredness and so on were considered to be relevant while tweets about broken bones and cancer were considered to be irrelevant. Also a lot of tweets depend on the history of the user, for example a tweet with the text "I feel so sick right now" will be considered as relevant but if we checked the history of the user this sickness might be related to cancer. This wasn't handled in the work as it'll require a more sophisticated system. Check Table II and Table III for more examples of the tweets.

TABLE I. THE POSITIONS OF THE STATES

| State | Latitude | Longitude | Radius |
|---|---|---|---|
| New York | 40.712776 | -74.005974 | 250Km |
| Washington | 47.751076 | -120.740135 | 250Km |
| New Jersey | 40.058323 | -74.405663 | 250Km |
| California | 36.778259 | -119.417931 | 250Km |
| Florida | 27.664827 | -81.515755 | 250Km |
| Louisiana | 30.984299 | -91.962334 | 250Km |
| Illinois | 40.633125 | -89.398529 | 250Km |
| Massachusetts | 42.407211 | -71.382439 | 250Km |

TABLE II. EXAMPLES OF TWEETS THAT BELONG TO THE TRUE CLASS

| Tweet | Label |
|---|---|
| if you have been going to frosty factory GO GET TESTED FOR CORONA 6 people have already tested positive from using the same microphone for karaoke. get tested or stay the hell away from me. | True |
| hehe... someone ah the blm protest has tested positive for corona ??????????????????gonna order a test tomorrow uhm | True |

TABLE III. EXAMPLES OF TWEETS THAT BELONG TO THE FALSE CLASS

| Tweet | Label |
|---|---|
| RT @Plat4omLive: common symptoms of COVID-19 include: fever dry cough tiredness Some Less common symptoms include: aches and pains sore throats | False |
| RT @HigherHealthSA: #MythBuster No. 4. Being able to hold your breath for prolonged periods of time does not mean you're #Covid19 free! The… | False |

### *E. Modeling and Evaluation*

This section basically explains how the first steps were taken into the prwork and resulted in the final implementation of the model. All the steps are previewed below alongside the tools and everything used to get to the results.

### *F. Initial Implementation of the Model*

To begin with, 200 samples of tweets were extracted worldwide not from a certain state or country. They were tested using the Naive Bayes classifier in RapidMiner, it was chosen as a result of its popularity in text classification. First of all the data was extracted from R and all the preprocessing has been applied, especially tokenization and Lowercase Tokens as the Naive Bayes classifier cannot work with strings directly it has to be converted to vectors that includes individual words. As for the model created in RapidMiner the input dataset which is the 200 tweets is split into two parts, 10% to be labelled and act as a training dataset and the other 90% acts as the testing dataset. When the data enters the model it'll output the classification prediction for the unlabeled data and the accuracy of the classifier. The output data was the text field which contained the text in the tweet as well as the label field which specifies whether someone is sick or has tested positive for Coronavirus or not, this field is binary either true or false. Other fields like Retweet count and the date were also present.

The first attempt resulted in a 100% accuracy, 100% class precision and a 100% class recall as shown in Table IV.

### *G. Continued Implementation of the Model*

Since the Naive Bayes classifier gave extraordinary results while working with a small dataset of 200 tweets, the next step was to try a larger dataset of 1000 tweets. The same criterion was followed in testing this amount of data except that the majority of the data were categorized to be False, i.e. indicating that people don't have Coronavirus nor the symptoms.

The results were great, the classifier gave an accuracy of 96.59% which is of course less accurate but still it's perfect. The rest of the results are present below in Table V.

To make sure that the right classifier was used, another popular classifier which is the SVM as it's also known to give decent results regarding text classification. The same experiment applied on the 200 tweets dataset on the Naive Bayes classifier was done once more but using the SVM as a classifier with the same dataset. The results were so far from good compared to the Naive Bayes classifier. SVM achieved 75% accuracy as shown in Table VI.

TABLE IV. PERFORMANCE VECTOR FOR 200 TWEETS USING NAIVE BAYES

| | True True | True False | Class Precision |
|---|---|---|---|
| Pred. True | 9 | 0 | 100% |
| Pred. False | 0 | 10 | 100% |
| Class recall | 100% | 100% | |

TABLE V. PERFORMANCE VECTOR FOR 1000 TWEETS USING NAIVE BAYES

| | True True | True False | Class Precision |
|---|---|---|---|
| Pred. False | 51 | 0 | 100% |
| Pred. True | 3 | 34 | 91.89% |
| Class recall | 94.4% | 100% | |

TABLE VI. PERFORMANCE VECTOR FOR 200 TWEETS USING SVM

| | True True | True False | Class Precision |
|---|---|---|---|
| Pred. False | 8 | 2 | 80% |
| Pred. True | 1 | 6 | 85.71% |
| Class recall | 88.89% | 75% | |

Based on the results, the rest of the work will be handled using the Naive Bayes classifier as it shows better accuracy results of 96.59% especially in classifying the minority class. Final implementation of the model.

The process executed in this phase was performed using the model in Fig 2. The first step is collecting the data from twitter from the 51 states in the United States over 7 days from the 18th of June 2020 till the 25th of June 2020 using the same keywords mentioned in the data collection section. The first step is to split the data into testing and training datasets. Data preprocessing is the next step to be executed in order to remove noisy data and convert it to a suitable format. Furthermore, performing building and testing of the model using Naive Bayes algorithms. Furthermore, the classification algorithm from RapidMiner is applied on the preprocessed data.

The training phase is very important, it plays a role in identifying patterns in the dataset in order to build an accurate model. A part of the classification training process one of the attributes should be assigned to act as a label column. This column is used to train the model and predict the positive tweets which is our aim. Basically, after selecting the label class and its value, it's used in calculating conditional probability based on the other attributes given in the same row.

After applying the model on the data from each State, the output should include the category of which the text is classified whether it's true or false as shown in Fig 3.

After the data is classified to their category it enters another model in order to output the word frequency in the total document, in the True class and in the False class. This will help us fig. out which words occur more in the True class which can be used later on as keywords in our search queries. Check Fig 4 and Table VII these are the words frequencies from one of the datasets of Ohio tweets.


Fig. 2. Classification Model.

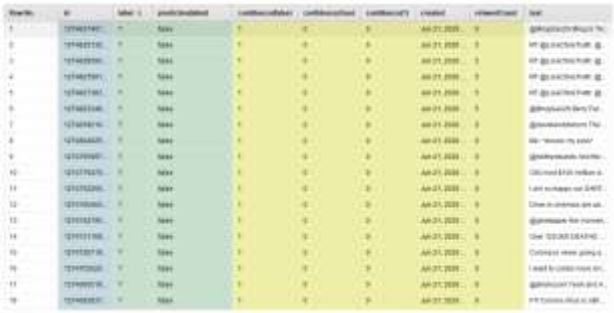

Fig. 3. Data Output in RapidMiner.


Fig. 4. Ohio Word List.

TABLE VII. TOP 5 WORDS IN THE DATASET OF OHIO

| Word | In documents | Total | Class (true) | Class (false) |
|---|---|---|---|---|
| Corona | 111 | 110 | 104 | 7 |
| Cough | 43 | 27 | 39 | 4 |
| Corona virus | 23 | 23 | 21 | 2 |
| Virus | 23 | 23 | 21 | 2 |
| Fever | 21 | 21 | 18 | 3 |

Another process added to the model is extracting the influence words from the class True based on multiple things. Retweet count, weight by correlation, weight by Gini index, weight by information gain and finally weight by information gain ratio as shown in Fig 5 The Retweet count is considered to be important in this case as the more retweets a tweet gets, the more important the tweets is. These different weights are then applied and their average is taken. Table VIII contains the top 3 influence words in the Ohio dataset.

*1) K. Sentiment analysis of tweets:* This part will include the explanation of the model that exists in Fig 6. Five-fold cross-validation was used, the main process was repeated 5 times but with different training and testing sets each time. This process was carried out in RapidMiner as shown in Fig 7. First, the data were split into two sets. 90% of the tweets were classified as the testing set, while the remaining 10% were for the training set. From the training data, all the n-grams were extracted. The training and testing datasets were converted into their corresponding feature vectors. A feature vector would always be 0 if the n-gram did not occur in the tweet. If it occurred in the tweet it will have the value of 1 (present) or the frequency, this depends on the configurations used. The training vectors were used to train a classifier. This classifier was the Naive Bayes, it would then classify the testing dataset into positive or negative. The label assigned to a tweet was compared to the actual class to see if they are the same. The statistics of the output data were saved for visualizations. The statistics of interest are the accuracy value, the confusion matrix and the classified tweets.

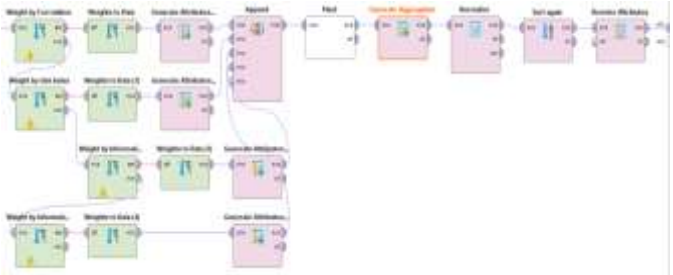

Fig. 5. Determining Influence in RapidMiner.

TABLE VIII. TOP 3 INFLUENCE WORDS IN OHIO DATASET

| Attribute | Importance |
|---|---|
| Fever | 1 |
| Corona | 0.934 |
| Cough | 0.924 |

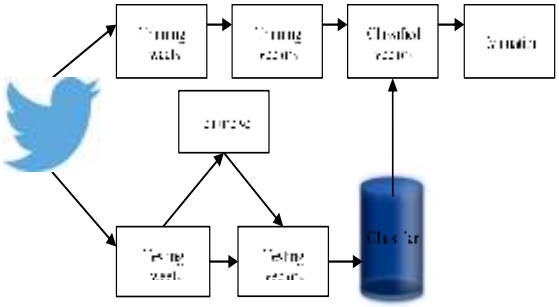

Fig. 6. Process of Sentiment Analysis.

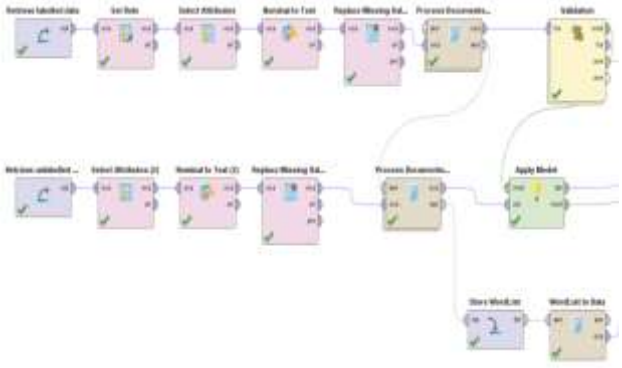

Fig. 7. Sentiment Process in RapidMiner.

*a) Deployment:* As mentioned previously in the business understanding phase, the objectives of this work is to extract tweets from Twitter related to the novel coronavirus and classify them based on True (sick) or False (not sick) then compare them with real world data and measure the correlation. Based on the results predictions were made and also sentiment analysis and people's consent were measured. In the model implementation, keywords that affected the classification were extracted in each State as well as statistical numbers in each State in America and visualizations will be discussed in the results section.

*b) Time series:* In order to evaluate the results and be able to compare it to real world data, it's better to see the data in a time series format. Two types of time series were used, the first was daily values and the second was weekly values.

To get a better understanding of the output data, the known number of positive tweets per day has to be recorded so it can be known if the data recorded was more or less than expected and use these numbers in many different applications. An example for one application, to support hospitals and their staff of how many cases per day to expect.

To model the time series, a package called the prophet was used in R to forecast. It's a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality.

## V. Results

The total positive tweets collected are 240,331 during the period of 17-06-2020 till 23-06-2020. This means that around 34,333 tweets were collected per day. After applying the Naive Bayes classifier on our data a plot chart was made to visualize the positive tweets gathered over each day as shown in Fig 8 alongside with the same plot chart but with the cumulative numbers in order to visualize the increases and decreases in the number of cases in Fig 9

After collecting data from each state in the United States. The positive numbers were visualized on the map in Fig 10. The darker the color gets on the map the higher the number.

As a result of having 51 states, the visualization for each one cannot be included in this paper but as an example positive numbers for both California and New York will be visualized and discussing their results in Fig 11 and Fig 12. In California no certain pattern was noticed over the days. The first day 17-06-2020 was a 4Wednesday and it had 4,108 positive cases then numbers kept increasing and decreasing but in the last two days they kept increasing with a high percentage to reach 5,545 and finally the peak which is 7,73. This pattern could be observed better if data for another week was extracted and be able to compare them and observe if they have the same pattern.

The New York numbers were a lot less than those of California. In total they were 4,898. It reached its peak on Friday the 19th by scoring 916 positive cases. It started decreasing till the last day to reach 627 positive cases. As stated that the numbers of the two states are very different from each other, they do not follow the same pattern so each State has to be visualized separately in order to get accurate results.

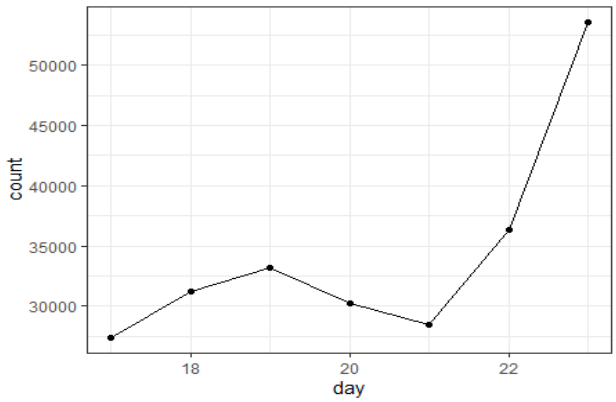


Fig. 8. Positive Tweets in the United States.

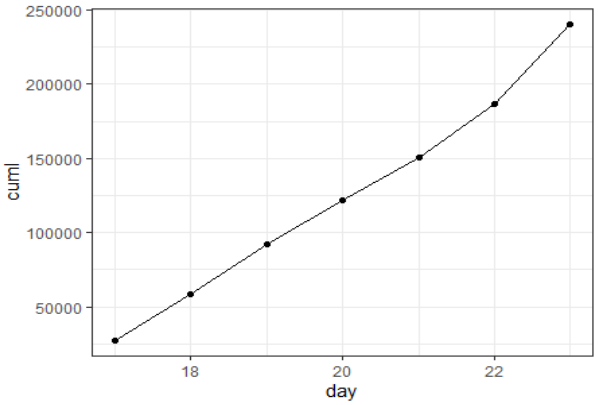


Fig. 9. Cumulative Positive Tweets in the United States.

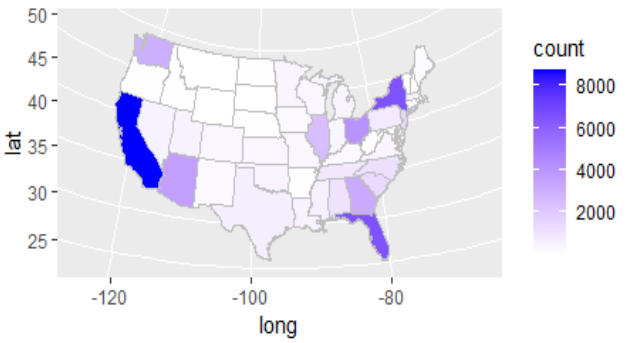


Fig. 10. Cases in Every State.

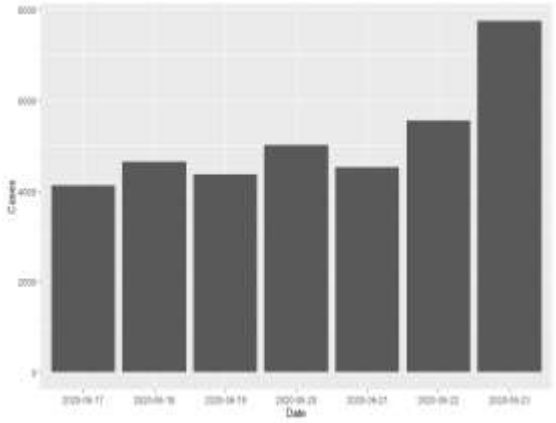

Fig. 11. Cases in California.

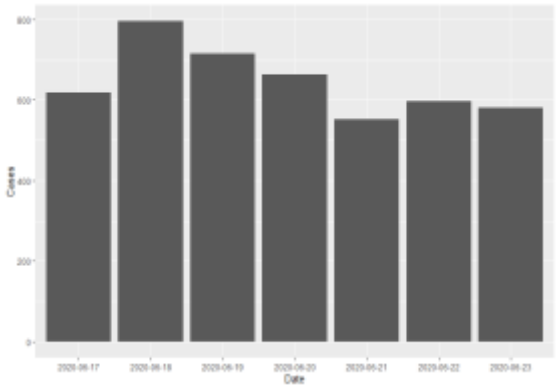

Fig. 12. Cases in New York.

## VI. Conclusion

In this paper, a model was designed in order to indicate if someone is sick or has tested positive for Coronavirus through twitter during a given day. To achieve this machine learning algorithms were used alongside statistics. The Naive Bayes classifier showed promising results in classifying the data accurately by achieving 96.59% accuracy after the preprocessing steps have been applied.

In order to be able to compare the classified tweets with real world data, Johns Hopkins dashboard was used to get the real data and compare them with each other. Both data were highly correlated.

The model also analyzed the word frequency to output the most frequent words that occurred in the True class and in the False class. Furthermore, influence words were extracted by calculating their weight by correlation, weight by Gini index and weight by information gain. The extracted influence words can therefore be used in search queries.

The performance for the predictions wasn't tested as mentioned before. The package used for this process was Prophet created in R-code. Hopefully, the performance of this predictor will give good results as our data is very close to being accurate.

The divergence between real cases and public sentiment could indicate many possible problems like public over- or underestimating of threats caused by diseases, because lack of the public media to reach and inform people properly to improve their awareness. It could also indicate lack of public healthcare infrastructures or deficiency of protection measures conducted by public authorities. Anyway, these new kinds of significant information can help decision maker in different application domains.

## Acknowledgment

The researchers extend their thanks and gratitude to the Deanship of Graduate Studies and Scientific Research at Dar Al Uloom University for their support and funding of this study.

Also, the researchers extend their thanks to the German University in Cairo (GUC).

## References

[1] C. Paquet, D, Coulombier, R. Kaiser, &N. Ciotti. (2006). Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill, 11(12), 212-214.

[2] M.Tsytsarau, & T. Palpanas, (2012). Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 24(3), 478-514.

[3] C. Tucker, & H. Kim, (2011). Predicting emerging product design trend by mining publicly available customer review data. In DS 68-6: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 6: Design Information and Knowledge, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011.

[4] T. Sakaki, M. Okazaki, & Y. Matsuo, (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860).

[5] A> P. Rajan, & S. Suresh, (2015). Application of Retail Analytics Using Association Rule Mining in Data Mining Techniques with Respect to Retail Supermarket. IJEMR, 5(1).

[6] N. Collier & S. Doan, (2011, November). Syndromic classification of twitter mesges. In International Conference on Electronic Healthcare (pp. 186-195). Springer, Berlin, Heidelberg.

[7] S.Tuarob & C. S. Tucker, (2013, August). Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data. In ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers Digital Collection.

[8] Van der Eijk, M., Faber, M. J., Aarts, J. W., Kremer, J. A., Munneke, M., & Bloem, B. R. (2013). Using online health communities to deliver patient-centered care to people with chronic conditions. Journal of medical Internet research, 15(6), e115.

[9] L. F. Lopes, J. Zamite, B. Tavares, F. Couto, F. Silva & M. J. Silva, (2009, September). Automated social network epidemic data collector. In INForum informatics symposium. Lisboa.

[10] B. W. Hesse, D. Hansen, T. Finholt, S. Munson, W. Kellogg & J. C. Thomas. (2010). Social participation in health 2.0. Computer, 43(11), 45-52.

[11] J. B. Long, & J. M. Ehrenfeld, (2020). The Role of Augmented Intelligence (AI) in Detecting and Preventing the Spread of Novel Coronavirus.

[12] Google flu trends: 'how does this work?' http://www.google .org/flutrends/about/how.html Accessed: 2-24-2020.

[13] L. C. Madoff, D. N. Fisman & T. Kass-Hout (2011). A new approach to monitoring dengue activity. PLoS neglected tropical diseases, 5(5). kasshout/edemocracy-egypts-18-day-revolution. Accessed: 2-24-2020.

[14] SentiWordNet. http://sentiwordnet.isti.cnr.it/. Accessed: 2-24-2020.

[15] S. J. Lee & K. Siau. (2001). A review of data mining techniques. Industrial Management & Data Systems.

[16] K. Wilson & J. S. Brownstein, (2009). Early detection of disease outbreaks using the Internet. Cmaj, 180(8), 829-831.

[17] M. Barstugan, U. Ozkaya & S. Ozturk. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. arXiv preprint arXiv:2003.09424.

[18] J. Wirth, R., & Hipp (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.