ARTIFICIAL INTELLIGENCE THEORY and APPLICATIONS

ISSN: 2757-9778 || **ISBN :** 978-605-69730-2-4

More information available at **aita.bakircay.edu.tr**

Detecting COVID-19 Pandemic Using Sentiment Analysis of Tweets

Syed Mujtaba HASSAN [1, *], Jawad KHAN [1], Muhammad Adnan KHAN [1], Muhammad Saeed KHAN [2]
Imran AHMAD[1], Mohsin KHAN [1]

[1] *Riphah International University, Riphah School of Computing and Innovation, Pakistan*
[2] *Lahore Leads University, Department of Electrical Engineering, Pakistan*

* Corresponding Author: Riphah International University Lahore, Riphah School of Computing and Innovation, Pakistan
Tel.: +92 3361719581. E-Mail: mujtaba.hassan@ripaha.edu.pk

Publication Information

Keywords :
- Twitter;
- COVID-19;
- Sentiment analysis;
- Machine learning;
- Data mining.

Category : Research article

Received : 07.07.2021
Accepted : 25.09.2021

A B S T R A C T

From 2019, the world is facing an unforeseen challenge in the form of COVID-19, which started in Wuhan (China), and within two months, it spread to 212 countries. The coronavirus disease (COVID-19) pandemic puts unprecedented pressure on healthcare systems worldwide. Due to its rapid widespread around the globe affecting the lives of millions, extensive measures to reduce and prevent its transmission have been implemented. One of which is to shut down their cities completely. During this Pandemic, people started to express their situations through social media tools. In natural language processing, valuable insights can be captured from textual data taken from different social media platforms. In this research work, data related to COVID-19 is collected from a popular social networking site, Twitter. The tweets gathered are refined through pre-processing for text mining and sentiment analysis. From this data, we successfully detect the actual count of people who may be affected by the COVID-19 Pandemic using sentimental analysis and machine learning techniques.

1. Introduction

In December 2019, the world was hit by a new challenge by the name of COVID-19. The pandemic starts in China and spread to 212 countries around the globe. Medical institutes are failing to stop the spread of the virus. Till this date, 6,033,875 people are affected by this, and 366,894 people lost their lives [1].

To slow the spread of the virus, governments enforced lockdown and curfew in their countries. Somehow the first wave of the virus is over. Healthcare institutes and policymakers need to prepare themselves for the second wave that is expected to be worse than the first one. Developing countries like Pakistan, India, Bangladesh, etc. are lagging in testing COVID-19 due to weak economies and less spending in the health sector. It is reported that the test ratio of over a thousand in countries like India, Pakistan, and Bangladesh is 2.71, 2.54, and 1.88 [3]. Which is lower than America, which has 51.17, Australia 57.04, and Russia 74.85. New techniques need to discover through which we can identify more cases without conducting tests [2].

Social media (SM) is becoming a fundamental need for everyone. More and more people are joining SM, as per the report, 2.95 billion people are using social media in 2019, which was 0.97 billion in 2010, and it expects to be 3.43 billion in 2023 [4]. SM is helping different parts of societies like health sciences, businesses, education, the entertainment industry. It is benefiting the business to improve their decision-making and market strategies. Companies can get better feedback and direct customer engagements that help the owners to improve their approach.

Same in the field of Health, we can analyze the benefits or drawbacks of some specific medicine among people. It can also be helping to predict the behavior of society for some diseases. Through SM, people are sharing their daily life activities, problems, recommendations, feedback, interests, and other personal stuff daily. Talking about Twitter, more than 300 million people are using this SM network, and each day 500 million tweets are sent. [5]

These tweets contain discussions related to politics, social issues, sports, the latest news, and personal activities. If we analyze the personal activities tweets, we can find a person's daily life problems and health conditions. Previously, a lot of work was performed by analyzing the user's social media profiles to access health conditions. This experiment was on a diabetic patient, which was helpful. So, we are proposing a similar system which needs to introduce in these conditions which will help the governments and health institutes to take a count of those people who were not tested and are affected by the same symptoms.

The aim of this research is the following:

- Recognizing and examine the data mining techniques through which we can extract the required information from the Social media network.
- Apply sentiment Analysis on the extracted data, which will help us to judge the condition and emotions of the user affected by this pandemic.

The goal is to detect COVID-19 pandemics by studying the tweets of the user, which use the words of symptoms in their tweets.

2. Literature Review

Our world faces a lot of diseases which infected a lot of people in the past. Our researchers study diseases like influenza and Ebola outbreak by applying different data mining algorithms. The results which they got from the above research can be beneficial for any upcoming conditions. Through the help of data mining algorithms, an influenza disease outbreak can be detected using SM, which are way faster than some health institute around the world.

Some of the research already made using tweets to predict the influenza spread rate in Portugal. The extracted the tweets which use the keyword like fever, flu, and other symptoms. Naive Bayes classifier applied to the tweets to determine and filter influenza-related data. In the end, linear regression is used to predict the future spread of this disease in society [6].

In China, research was done on two diseases outbreak, one in 2012, which was the MERS-COV, and H7N9, which outbreak in 2013. Researchers use the Weibo website to collect the reaction of people to the above two diseases. They filter the users who have followers of more than 10000, which makes them credible. The researcher makes a list of symptoms and uses these keywords to collect the post, which uses these keywords. The goal of this research is to see the reaction of Chinese users to these two outbreaks [7].

Similar research made by Culotta is to collect half a billion tweets that contain some information related to influenza rate and alcohol sales. In the first step, Culotta uses a document classifier to filter the useless data. In the end, the researcher implements SVM, Decision tree, and logistic regression for classifying the tweets [8]. In 2009 a surveillance system was designed which was capable of collecting more than 2 million tweets of users infected by Influenza A virus subtype (H1N1) diseases through keywords like swine flu, Fever, Chills, Cough, and Sore throat. This system identifies the users who use these keywords in their tweets that can affect an H1N1 patient [9].

In 2018 another research was made to detect two infectious diseases Ebola and sine flue through collecting tweets from Twitter based on the keyword's flue, H1N1, and swine flu [10].

One hundred fifty-nine thousand eight hundred two tweets from the eleven USA cities users collected contain the keyword flu in them. Different machine learning algorithms are implemented on them to classify the tweets and divide them into valid or invalid. The valid is the users who have flu symptoms based on the keywords found in their tweets [11].

A lot of work has already been done to detect different diseases and implement machine learning algorithms, to classify the data and produce better results from it. These will help us find a new way to support our research in finding solutions to detect COVID-19 patients with the help of SM.

3. Problem Statement

This research facilitates the countries of the world, spending a tiny amount of money on their health sectors and failing to conduct COVID-19 tests as compared to the developed counties. This research will help the policymakers and health institutes to detect patients who are suffering from similar COVID-19 symptoms.

4. Methodology

As we discuss the research problem, we need to find ways to get useful information from social media networks through which we can access the people affected by COVID-19. For this, we choose Twitter as our source of data. We collected the data containing information about the user name, location, tweet content, and engagement of his status. As our research location, we chose South Asian countries like Pakistan, India, and Bangladesh.

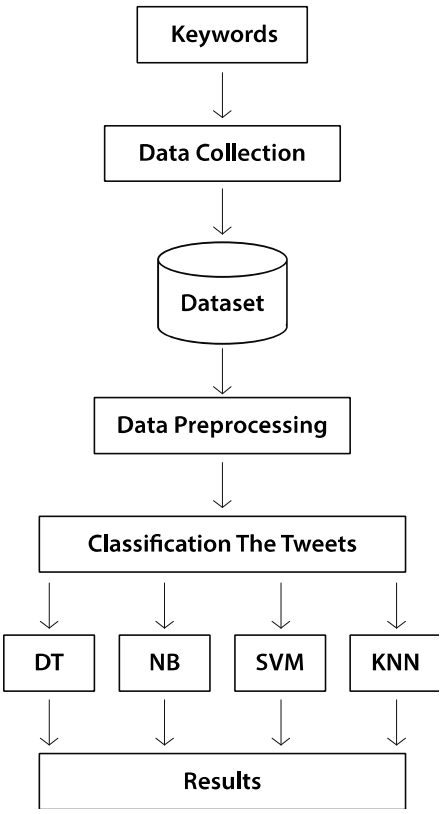

FIG. 1. FLOW CHART OF THE PROPOSED SYSTEM

Figure 1 shows the proposed system flowchart diagram.

The tool which we are using throughout this research is Orange, which is an open-source machine learning and data visualization tool. We applied some preprocessing techniques to make our data more useful and remove the useless data from it. We used preprocessing methods in our study, which were tokenization, n-grams, filter stop words, and stemming. This research divides the collected data into two categories one is valid, and the other is invalid. For evaluating our research, we will compare data mining techniques and apply an accuracy test. The data mining algorithms which we used in this research are the following: Support Vector Machine (SVM), Naive Bayes (NB), K-nearest neighbor (k-NN), and Decision Tree (DT).

We can see the flow chart of the process in Figure 1. This process helps us to detect COVID-19 through tweets using keywords related to its symptoms.

4.1 Data collection

To conduct our research, we use Twitter as a data source. For collecting data from Twitter, a Twitter development account is needed, which provides you with an API key and an API security key. We add to our Orange tool through which we access the data from it. We are collecting data from countries in South Asia, India, Pakistan, and Bangladesh. We received 10000 Tweets from these countries, which are using these thirty-four keywords some of which are following.

- COVID-19
- Suffering
- Fever
- Dry cough
- Tiredness
- Aches and pains
- Sore throat
- Diarrhea
- Conjunctivitis
- Headache
- Loss of taste or smell
- Rash on Skin

Based on these tweets using the above keywords, we classify the tweets into "valid" and "invalid" data. The valid tweets are the one which processes the symptoms of COVID-19 in them; for example, I am suffering from fever and dry cough. In this statement, the author of these tweets is suffering from two of the primary symptoms of COVID-19. So, the tweets like these will be classified into a group of valid tweets. The tweets in the invalid category are those which are not related to the COVID-19 pandemic, and the topic discussed is out of context.

For selecting the keywords, we use different sites like google and WHO sites to get the users searching history through this time of COVID-19. First, we take the keyword, which was related to COVID-19. Then we check the trends that people are searching for from google in this time of the pandemic. By doing this, the number of tweets that we extracted from Twitter using their API was reduced to 840. This first step is done in orange. The second step was to analyze the data manually and then remove them from the sheet. After the second step, the tweets we got after performing the first step were 840 and reduced to 467.

4.2 Preprocessing

Before going to our experiment, there is an essential step that requires to be done. This step is known as preprocessing. In this step, we arrange and process the data according to our experimental requirement.

There are a lot of sub-steps in this preprocessing through which we can get the expected results. Some of the steps which we used in our research are the following.

Filtering:

In this process, the goal is to remove the undesired data that exist in our data set. We can do filtering by the following steps.

- Stop words
- Lexicon
- Regexp
- Document Frequency
- Most Frequent Tokens

The most essential and universal filtering step is the stop word. It is being used in all the sentiment analysis experiments. The primary job of stop words is to identify the words which the document doesn't need. We provide a list of such Keywords to the tool, and whenever the tool finds any such keyword, it stops and removes that specific data from the data set. Lexicon is another useful filtering technique that helps us to check the words or vocabulary of a language. We can use this part to filter the text document on a lexicon basis. Regexp helps us to remove the symbols we use in our text like dot, comma, semicolon, etc. Regexp removes the punctuation mark and converts the document as the only word from it so we can analyze the keywords and collect them in some bags. Document Frequency helps the Researcher to find the importance of the specific word in the document. It tells the user that if we remove a particular keyword, the meaning of the sentence will be changed.

Tokenization:

This is a process through which we divide the sentences into smaller chunks, which can be words or phrases. The reason to do that is to analyze the data at its lowest form. Each word is assigned a number, which is its primary key.

N-Grams:

To help a machine to understand the word in a context to get a better understanding. The working of N-grams is like this; it checks the name before the keyword and the word after it and then judges the meaning of that keyword in that sentence. We can set the range of the word to compare with, like in our case, let's take an example. I am suffering from a high fever. Now when we put this statement in the N-grams, and we make the range of comparing 3, then this will do like "I am suffering" next will be "suffering from high" then "with high fever" So we got three answers from this which the machine can understand. One is how suffering is? Second is suffering with? And third is the intensity of pain? That's how N-grams works.

Normalization:

The last step after which the research can start the experimental part is Normalization. Normalization is a technique in which the derived words are converted back to their original form. Normalization helps the machine to understand the basics of the document. It also makes the machine comfortable to get a better understanding of the things which the text wants to convey. In Orange, Normalization is further divided into four parts: Porter Stemmer, Snow Stemmer, WordNet Lemmatizer, and UDPipe Lemmatizer.

4.3 Classififcation

Classification is a process in Machine learning through the Machine that can categories the data into a group. This process is done by dividing the data into a training set and testing set. The training set helps the Machine to judge the data to which category it will fit by analyzing the properties of the data. The results which we got from the training data we apply this to our testing data and see what the outcomes will be.

We used three classification algorithms in our research, which are Decision Tree Algorithm (DT), *Naïve* Bayes Algorithm (NB), Support Vector Machine Algorithm (SVM). We used different parameters for them for training and testing. The working of these algorithms is discussed in detail below.

Decision Tree Algorithm (DT)

The Decision Tree algorithm is a machine learning algorithm that is used for solving classification and regression problems. The Decision Tree algorithm is a member of supervised learning. The primary goal of the decision tree is to predict the value with the help of the previous data we obtained from the training data methods. The process which we follow in the decision tree is to predict the labels. We compare the record attribute with the root attribute. Once we analyze the root values, we move the next node, and that's how it works, we follow this and go from the bottom to up.

Naïve Bayes Algorithm (NB)

Naïve Bayes Algorithm is the easiest and fastest classification algorithm that has the capability of predicting data in real time. This algorithm is based on the Bayes Theorem. In Bayes Theorem, every product feature that depended on each other also contribute their properties independently. We can take these properties and predict them. The equation of the Bayes theorem is the following.

$$P(C|X) = P(X|C)\, P(C) / P(X)$$

- $P(C|X)$: the posterior probability of *class* (C, *target*) given *predictor* (X, *attributes*).
- $P(C)$: the prior probability of *class*.
- $P(X|C)$: the likelihood, which is the probability of the *predictor* given *class*.
- $P(X)$: the prior probability of *predictor*.

Support Vector Machine Algorithm (SVM)

Another machine learning algorithm we are using in our research is Support Vector Machine (SVM), a supervised learning algorithm. We can use SVM both for regression as well as classification purposes. The working of the SVM algorithm is, we plot all the data as a point in n-dimensional space where data of the same type is in the same group.

K-Nearest Neighbor Algorithm (k-NN)

K-Nearest Neighbor Algorithm (*k-NN*) is another supervised machine learning algorithm that can perform both regression and classification. In *k*-NN the algorithm works in finding the distance between the training data and the test data on which classification is applied. The primary purpose of using this algorithm is to detect the similarity among the different texts.

5. Results and Discussion

In this research, the first step was the preprocessing techniques we used to clean our data and make the data ready to use. We choose four classification algorithms with which we will produce the results. The algorithms which we used are; Decision Tree Algorithm (DT), *Naïve* Bayes Algorithm (NB), Support Vector

Machine Algorithm (SVM). We divide the dataset into a training dataset and a testing dataset based on the k-fold values. We apply different fold values to our values. The fold value which we use is 5,10 and 20. To evaluate our results, we have to apply some metrics to calculate the performance of our algorithms. These are Accuracy, F1-measure, Precision, and Recall. The formula of these are the following:

Accuracy = (TP + TN)/(TP + TN + FP + FN)

Precision =TP/(TP +FP)

Recall = TP/(TP + FN)

F 1 − measure = 2 ∗ ((Precision ∗ Recall)/(Precision + Recall))

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).

We conducted our experiment on the tweets which are related to COVID-19 by using multiple folds values.

We used multiple folds that have the values 5,10 and 20. The tweets are then tested with different folds and different classifiers such as NB, DT, SVM, and k-NN. Table 1, Table 2, and Table 3 display the values which we collected as a result of this experiment. To verify our analysis, we apply accuracy, precision, recall, and F1-measure test. The figure shows the graphical representation of the research.

TABLE 1. EVALUATION RESULTS WITH 5 NUMBER OF FOLDS

Models	***Accuracy***	***F1−measure***	***Precision***	***Recall***
k-NN	0.581	0.269	0.767	0.246
Decision Tree	0.637	0.701	0.690	0.676
Support Vector Machine	0.818	0.673	0.596	0.772
Naïve Bayes	0.516	0.000	0.000	0.000

TABLE 2. EVALUATION RESULTS WITH 10 NUMBER OF FOLDS

Models	***Accuracy***	***F1−measure***	***Precision***	***Recall***
k-NN	0.597	0.259	0.778	0.246
Decision Tree	0.615	0.674	0.672	0.676
Support Vector Machine	0.809	0.673	0.596	0.772
Naïve Bayes	0.516	0.004	0.120	0.002

TABLE 3. EVALUATION RESULTS WITH 20 NUMBER OF FOLDS

Models	*Accuracy*	*F1−measure*	*Precision*	*Recall*
k-NN	0.605	0.238	0.778	0.246
Decision Tree	0.682	0.735	0.730	0.742
Support Vector Machine	0.811	0.673	0.596	0.772
Naïve Bayes	0.516	0.004	0.120	0.002

The evaluation results of the prosed study with multiple performance measure parameters as shown in Tables 1, 2, and 3.

6. Conclusion and Evaluation

In this paper, we discussed an approach to detect COVID-19 pandemic using sentiment analysis of tweets. For this purpose, we chose Twitter as a data source and collected a vast number of tweets in which keywords related to COVID-19 were detected. After receiving this data, we applied different preprocessing techniques to filter the data and divide it into two labels. One is valid, and the other is invalid.

After preprocessing, we used classification algorithms to identify a set of categories of the data it belongs to. For classification, we used the Decision Tree Algorithm (DT), Naïve Bayes Algorithm (NB) and Support Vector Machine Algorithm (SVM). We used four different k-fold values to determine the best suitable way for our experiments. This step was followed by a performance evaluation step, in which the research applied accuracy tests. From experiments, we have found that SVM best fits for classification of our data in terms of accuracy. This research shows that we can use this type of approach to detect patients suffering from COVID-19 and are not tested for any reason.

7. Future Work

In the future, we are planning to use other social media networks to get more data. This research also needs to be tested with more deep learning techniques so we can get better accuracy results.

References

[1] Joe Hasell, E. M., Diana Beltekian, Bobbie Macdonald, Charlie Giattino, Esteban Ortiz-Ospina, Hannah Ritchie, and Max Roser (2020). "Coronavirus (COVID-19) Testing." from https://ourworldindata.org/coronavirus-testing.

[2] Worldometer (2020). "Worldometer COVID-19 Data." from https://www.worldometers.info/coronavirus/.

[3] Joe Hasell, E. M., Diana Beltekian, Bobbie Macdonald, Charlie Giattino, Esteban Ortiz-Ospina, Hannah Ritchie, and Max Roser (2020). "Coronavirus (COVID-19) Testing." from https://ourworldindata.org/coronavirus-testing.

[4] Clement, J. (2020). "Number of social network users worldwide from 2010 to 2023." from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[5] Brew, S. (2020). "100 Social Media Statistics For Marketers in 2020 ". from https://statusbrew.com/insights/social-media-statistics-2020/.

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

[7] 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.

[8] 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

[9] 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..

[10] 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.

[11] 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.