Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14
© 2022 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

USING MACHINE LEARNING TO PREDICT THE SENTIMENT OF ARABIC TWEETS RELATED TO COVID-19

ESLAM AL MAGHAYREH[1,2]

[1] College of Applied Computer Science, King Saud University, KSA

[2] Computer Science Department, Yarmouk University, Jordan

E-mail: [1] ealmaghayreh@ksu.edu.sa, [2] eslam@yu.edu.jo

ABSTRACT

The world has suffered a lot due to the spread of the COVID-19 virus. The world health organization (WHO) declared the COVID-19 pandemic a global emergency. The governments and healthcare officials everywhere were fighting to control the spread of this pandemic. Meanwhile, a considerable amount of social media data (e.g., tweets) related to COVID-19 is being generated continuously. In this paper, we will build a model that can identify the sentiment of Twitter data related to COVID-19 using machine learning. We will focus on analyzing Arabic language tweets to determine people's opinions, feelings, and status on the impact of COVID-19. The main objective of this research is to support efforts to study the impact of the COVID-19 pandemic on society. To achieve this objective, we have prepared a dataset of Arabic tweets related to COVID-19 and manually classified the tweets in the dataset. Then we have used machine learning to develop an approach to assess people's feelings about COVID-19. This approach can help the government and healthcare officials to identify any negative and positive aspects of this crisis to improve their response to similar future crises.

Keywords: *Sentiment Analysis, Machine Learning, Text Analysis, Natural Language Processing, Data Science*

1. INTRODUCTION

We are living in a world drowning in data. A tremendous amount of data is being generated every second. Social media platforms like Twitter and Facebook generate vast text data daily. Consequently, data science and big data analytics techniques have recently attracted significant attention. Data scientists use these techniques to analyze large data sets to extract insights and valuable information that could benefit several organizations.

Data science aims to collect data, explore it, understand it, and extract valuable knowledge and insights from structured and unstructured data to improve different aspects of our everyday lives. Facebook, Twitter, and many other social media sites continuously generate data. Customers are writing reviews for products on many e-commerce sites. Sensors of different types are collecting data from different sources worldwide. A smartphone builds up a record of your location and speed every second of every day.

Machine learning has been extensively exploited in the data science domain. Machine learning is a field of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Any successful data scientist must be familiar with different machine learning algorithms. R and Python languages (two of the most commonly used languages in data science) have packages implementing the well-known machine learning algorithms.

Natural Language Processing (NLP) is currently applied in data science research. A well-known application of NLP is sentiment analysis. Thousands of text documents can be processed for the sentiment in seconds, compared to days it would take a team of people to complete the same task manually. Sentiment analysis has applications in several domains. For example, companies may use sentiment analysis to help them evaluate their products.

COVID-19 has impacted almost every sector of life like the economy, education, religion, tourism, sports, etc. This paper is dedicated to exploiting data science and artificial intelligence techniques to support the efforts spent to study the impact of the COVID-19 pandemic on society globally in general and more specifically in regions speaking the Arabic language.

The main contributions of this paper are:

1. Preparing a COVID-19 dataset in Arabic from Twitter for sentiment analysis and opinion mining.

2. Exploiting machine learning to build sentiment analysis and opinion mining techniques to get as many insights out of the COVID-19 dataset as possible. The extracted insights could be beneficial in studying and controlling the impact of the COVID-19 pandemic on society globally in general and more specifically in regions speaking Arabic.

Data science and big data analytics techniques could be employed to study people's reactions to the pandemic and the steps adopted to control the spread of the pandemic. This could help the officials in assessing the usefulness of these steps.

The COVID-19 dataset prepared in this paper will significantly benefit other researchers working in data science and big data analytics. The results in this paper will help in identifying the most appropriate data science and artificial intelligence techniques to be used to extract insights out of text-based datasets.

In our research project, we will go through the following phases:

1. Data collection: in this phase, we will develop programs to collect the data related to COVID-19 from Twitter.

2. Data cleaning and preprocessing: The data collected from social media is not ready for analysis. A sequence of preprocessing steps has to be applied to the data to clean it and make it ready for analysis. Common examples of preprocessing steps are stopwords removal and stemming.

3. Model building and evaluation: we will explore the dataset prepared in the first two phases and identify the suitable data science and artificial intelligence techniques to extract insights from the data. We will also evaluate the performance of these techniques.

Figure 1 depicts the framework of the proposed sentiment analysis model for Arabic tweets related to COVID-19.

The remaining part of this paper is organized as follows: Section 2 presents some of the essential related works. Section 3 presents the main steps we have gone through to prepare the dataset used in this paper. In Section 4, we have presented the features extraction and model building steps. To evaluate the models built, we have conducted a set of experiments. The results of these experiments are presented in Section 5. Finally, the conclusions and future work directions are presented in Section 6.

2. RELATED WORK

Data science and artificial intelligence techniques could be employed to analyze social media data to extract insights valuable for organizations in several domains [1, 2, 3, 4]. Several researchers have employed these techniques to analyze data related to several pandemics to extract insights that can help stop the spread of these pandemics in the future [5, 6, 7].

Many research papers focus on exploiting data science and artificial intelligence to analyze data related to the flu pandemic. In [8], V. Lampos and N. Cristianini proposed the Social Network Enabled Flu Trends (SNEFT) framework, which monitors messages posted on Twitter with a mention of flu indicators to track and predict the emergence and spread of an influenza epidemic in a population. In [9], a machine learning algorithm has been proposed as a practical computational system to support fast and effective decisions in epochs when a flu virus begins spreading in a large or middle-size city.

Several researchers in the domain of data science and big data analytics adopt social media as the primary data source in their studies. In [10], the authors present an investigation into how online resources related to Swine Flu were discussed on Twitter. They have focused on identifying and analyzing the popularity of trusted information sources (i.e., official health agencies). They found that reputable sources are more popular than untrusted ones. However, they found that untrusted can still leak into the network and potentially cause harm. In [11], Hong, Yang, and Sinnott, Richard O. addressed the problem of disease outbreak detection via an automated and scalable cloud-based system for collecting, tracking, and analyzing social media data.

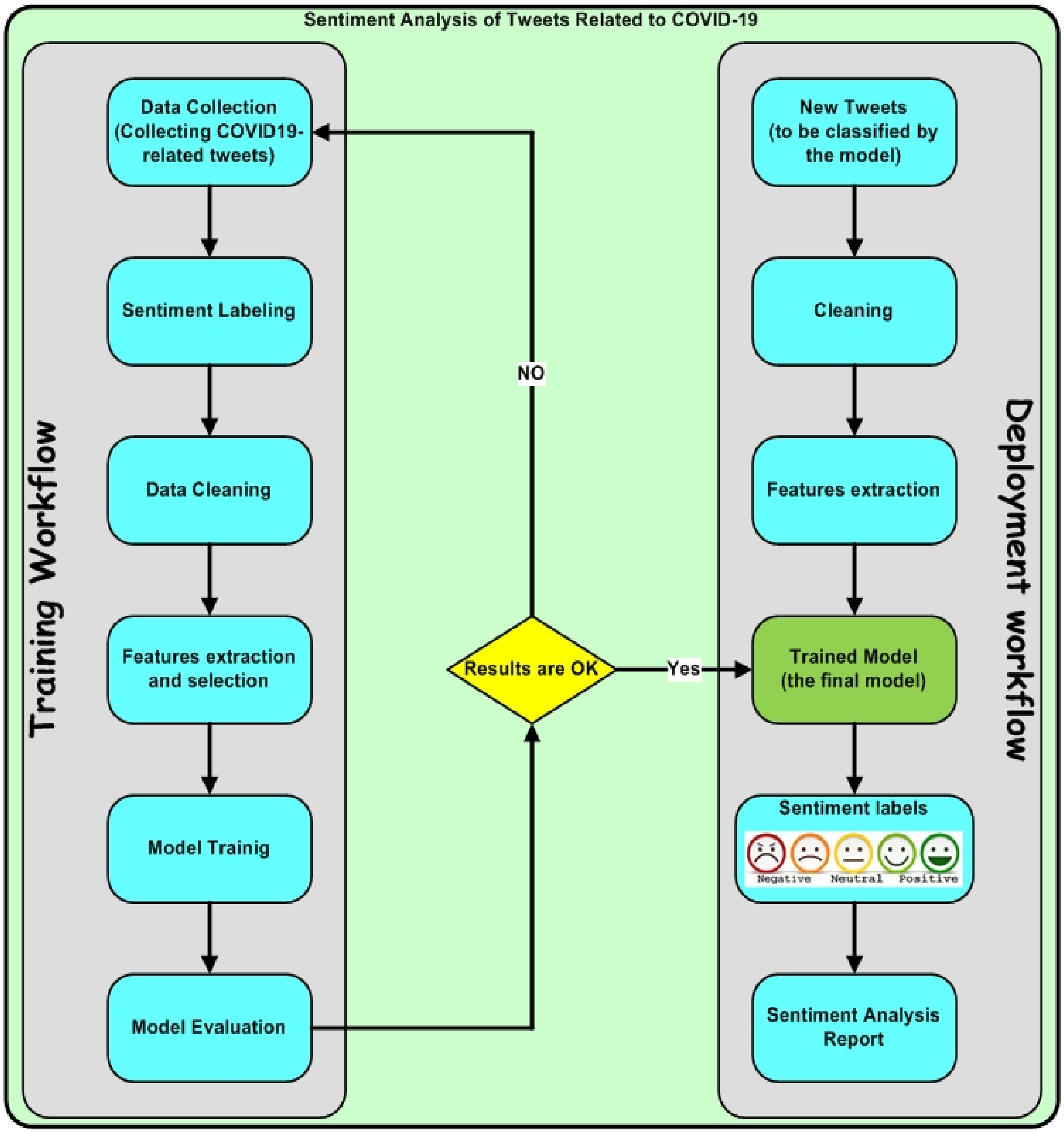

Sentiment Analysis of Tweets Related to COVID-19
Training Workflow
Data Collection (Collecting COVID19-related tweets)
Sentiment Labeling
Data Cleaning
Features extraction and selection
Model Trainig
Model Evaluation
NO
Results are OK
Yes
Deployment workflow
New Tweets (to be classified by the model)
Cleaning
Features extraction
Trained Model (the final model)
Sentiment labels
Negative
Neutral
Positive
Sentiment Analysis Report

Figure 1 The Framework of the proposed model for sentiment Analysis of COVID19 Arabic tweets.

K. Lee et al. have presented a model that predicts the weekly percentage of the U.S. population with Influenza-Like Illness using a multilayer perceptron with a back propagation algorithm on a large-scale social media stream [12]. T. Tran and K. Lee [13], have conducted a comprehensive study of understanding and mining the spread of Ebola-related information on social media.

Recently, researchers have started to exploit data science and big data analytics techniques to help in addressing several issues related to the COVID-19 pandemic [14]. In [15], the authors have proposed an artificial intelligence framework with predictive analytics capabilities applied to actual COVID-19 patients' data to provide efficient clinical decision-making support. Li et al. [16] exploit natural language processing techniques to classify the COVID-19-related information into seven types of situational information. Their work is based on data collected from Weibo (A Chinese microblogging website. Launched by Sina Corporation on 14 August 2009, one of the biggest social media platforms in China). In [17], the authors evaluated the influence of information (formal and informal) sources on public situational awareness for adopting health-protective behaviors such as social distancing.

In the following section of the paper, we will describe the details of the main phase of our research project, namely, data preparation.

3. DATA PREPARATION

Dataset preparation is a difficult task in any data science project. It is considered the most time-consuming phase of the data analytics life cycle. The first thing to be done in this phase is to collect as much data as possible from several sources. However, to use the collected data, we need to transform the raw data collected into a form fit for use with machine learning algorithms. This process is referred to as data preprocessing and cleaning.

Data preprocessing and cleaning involve exploring the data to see if any problems make it inappropriate for use in subsequent project steps. Problems in data could be duplicates, incorrect data, missing values, and data in the wrong format, etc.

We will focus in this paper on the sentiment analysis of Arabic tweets related to the COVID-19 pandemic. Consequently, we have developed programs to collect data related to COVID-19 from Twitter. We have collected around 10000 tweets in total. Figure 2 shows a sample of the assembled tweets.

Usually, the data collected from social media is not ready for analysis. A sequence of preprocessing steps has to be applied to the data to clean it (removing data not related to COVID-19, removing duplicates, removing punctuation, etc.) and make it ready for analysis. Common examples of preprocessing steps are stopwords removal and stemming.

We need to prepare a high-quality training dataset before we can apply machine learning algorithms to get insights out of the assembled data. This dataset must be labeled manually by a human. This is how we can train the machine learning algorithms so that they can perform tasks usually performed by intelligent people only. The performance of the machine learning algorithms relies on the quality of the prepared training dataset. Consequently, we have assigned the same data to several assistants to ensure the quality of labeling. Each tweet has been assigned a label indicating its sentiment (positive, negative, and neutral). The prepared dataset is one of the contributions of this paper, and it will be used as the training dataset for machine learning algorithms that will be used in the subsequent phases of this paper. Moreover, the dataset can be used by other projects related to sentiment analysis of Arabic text.

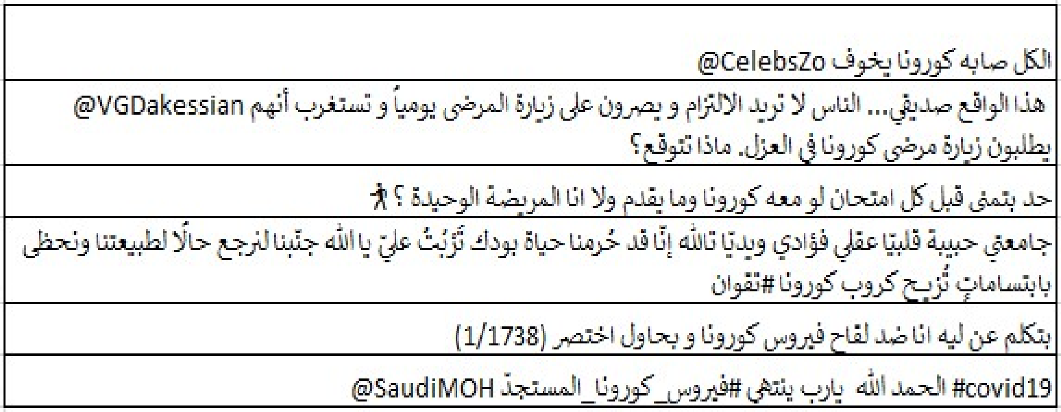
After cleaning, preprocessing, and labeling, we have 6903 labeled tweets in the final dataset. The resulting labeled dataset will then be used as input to machine learning algorithms that can automatically identify the sentiment of tweets collected in the future.

The following section will describe the features extraction and model building phases.

4. FEATURES EXTRACTION AND MODEL BUILDING

Once the dataset is ready, we can directly move to the second critical phase of our project. This phase aims to exploit machine learning algorithms to build a model to predict the sentiment of tweets related to COVID-19. However, machine learning algorithms cannot deal with texts directly. Consequently, an important step has to be performed before building the machine learning models, namely, feature extraction.

In the feature extractions step, we have used well- known techniques to extract numerical features from the textual data we have in hand. Initially, we have extracted basic numerical features like number of words, length of tweets, average word length, number of hashtags, etc.



الكل صابه كورونا يخوف @CelebsZo
هذا الواقع صديقي... الناس لا تريد الالتزام و يصرون على زيارة المرضى يومياً و تستغرب أنهم @VGDakessian يطلبون زيارة مرضى كورونا في العزل. ماذا تتوقع؟
حد بتمنى قبل كل امتحان لو معه كورونا وما يقدم ولا انا المريضة الوحيدة ؟ 🚶
جامعتي حبيبة قلبيّا عقلي فؤادي وبديّا تالله إنّا قد حُرمنا حياة بودك تُزِيْتُ عليّ يا الله جنّبنا لنرجع حالًا لطبيعتنا ونحظى بابتساماتٍ تُزيح كروب كورونا #تقوان
بتكلم عن ليه انا ضد لقاح فيروس كورونا و بحاول اختصر (1/1738)
#covid19 الحمد الله يارب ينتهي #فيروس_كورونا_المستجدّ @SaudiMOH

Figure 2 Sample of the collected tweets.

We have also used advanced techniques to extract numerical features like Bag Of Words (BOW) and term-frequency inverse-document-frequency (tf-idf) techniques. Fortunately, python libraries involve several functions to simplify the process of feature extraction. Once the features are appropriately extracted, we can move to the next step, namely, building the machine learning model to predict the sentiment of a tweet. We will first describe the machine learning models we have used to accomplish this task.

1. Naive Bayes (NB) is an algorithm that uses Bayes Theorem to classify objects and predict the probability of different classes based on different attributes. Typical uses of naive Bayes classifiers include medical diagnosis, spam filters, and text analysis. Naive Bayes is also known as Simple Bayes or Independence Bayes.
2. Decision Tree (DT) Algorithm is a supervised machine learning algorithm. It can be used for both classification problems and regression problems. This algorithm aims to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal nodes of the tree.
3. Random Forest (RF) is a supervised machine learning algorithm. It is an algorithm used for classification and is made up of many decision trees, and its prediction is more accurate than any individual tree.
4. Logistic Regression (LR) is a supervised classification algorithm. It predicts the probability of occurrence of an event by fitting data to a logit function.

5. EXPERIMENTAL RESULTS

This section will present the results of several experiments conducted to select the best possible model to predict the sentiment of an Arabic tweet related to COVID-19. We have used several feature extraction techniques and machine learning algorithms to build several prediction models. The purpose of the experiments is to identify the best prediction model. We have evaluated the models in terms of accuracy, precision, recall, and f1-score.

In the first set of experiments, we used the term-frequency inverse-document-frequency (tf-idf) method for feature extraction. Table 1 summarizes the performance of four different models to predict the polarity of an Arabic tweet (positive, negative). The results show that models based on the NB, LR, and RF machine learning algorithms are the best in terms of accuracy (0.76). The models based on the NB and LR algorithms are the best in terms of precision (0.76). The model based on RF is the best in terms of recall (0.9), and the models based on LR and RF algorithms are the best in terms of f1-score (0.82).

In the second set of experiments, we have used the bag of words (BOW) method for feature extraction. Table 2 summarizes the performance of four different models to predict the polarity of an Arabic tweet (positive, negative). The results show that the RF machine learning algorithm model is the best in terms of accuracy (0.77), precision (0.76), recall (0.88), and f1- score (0.82).

In the third set of experiments, we have used an algorithm to convert each Arabic word that appeared in the dataset to its root, and we have used the term-frequency inverse-document-frequency (tf-idf) method for features extraction. Table 3 summarizes the performance of four different models to predict the polarity of an Arabic tweet (positive, negative). The results show that models based on the NB and RF machine learning algorithms are the best in terms of accuracy (0.76). The model based on the NB algorithm is the best in terms of precision (0.74), and the model based on the RF machine learning algorithm is the best in terms of recall (0.92) and f1-score (0.82).

In the fourth set of experiments, we used an algorithm to convert each Arabic word that appeared in the dataset to its root and used the bag of words (BOW) method for features extraction. Table 4 summarizes the performance of four different models to predict the polarity of an Arabic tweet (positive, negative). The results show that the NB machine learning algorithm model is the best in terms of accuracy (0.78), precision (0.78), recall (0.87), and f1-score (0.82).

6. CONCLUSIONS AND FUTURE WORKS

In this paper, we have built several machine learning models to predict the sentiment of an Arabic tweet related to COVID-19. The results of the work in this paper are helpful for organizations working on studying the impact of the COVID-19 pandemic on several aspects of our everyday lives. We have built a sentiment data set consisting of thousands of labeled Arabic tweets. The experiments presented in this paper show that the sentiment prediction model based on the Nave Bayes machine learning algorithm with BOW features has the highest accuracy (0.78) and the highest precision (0.78). The model based on the Random forest algorithm with tf-idf features and stemming has the highest recall (0.92) and f1-score (0.82).

We can extend the work presented in this paper by considering other machine learning algorithms and deep learning in building the sentiment prediction model. We can also use word embedding to improve the feature extraction step. Finally, we can extend the dataset prepared by collecting and labeling more Arabic tweets related to COVID-19.

REFERENCES:

[1] Hanane Elfaik and El Habib Nfaoui. Deep Attentional Bidirectional LSTM for Arabic Sentiment Analysis in Twitter. In 2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA), pages 1–8, 2021.

[2] Faxi Yuan and Rui Liu. Mining Social Media Data for Rapid Damage Assessment during Hurricane Matthew: Feasibility Study. Journal of Computing in Civil Engineering, 34(3), May 1, 2020.

[3] Hyo Jin Do, Chae-Gyun Lim, You Jin Kim, and Ho-Jin Choi. Analyzing emotions in twitter during a crisis: A case study of the 2015 Middle East Respiratory Syndrome outbreak in Korea. In 2016 International Conference on Big Data and Smart Computing (BigComp), pages 415–418, 2016.

[4] Mohammad Tufail Malik, Abba Gumel, Laura H. Thompson, Trevor Strome, and Salaheddin M. Mahmud. Google Flu Trends and Emergency Department Triage Data Predicted the 2009 Pandemic H1N1 Waves in Manitoba. Canadian Journal of Public Health-revue Canadienne De Sante Publique, 102(4):294–297, JUL-AUG, 2011.

[5] Tsung-Hau Chen, Yung-Chiao Chen, Jiann-Liang Chen, and Fu-Chi Chang. Flu Trend Prediction Based on Massive Data Analysis. In 3rd IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, pages 304–308, 2018.

[6] Steven J. Hoffman and Victoria Justicz. Automatically quantifying the scientific quality and sensationalism of news records mentioning pandemics: validating a maximum entropy machine-learning model. Journal of Clinical Epidemiology, 75:47–55, JUL 2016.

[7] D. Kim, S. Hong, S. Choi, and T. Yoon. Analysis of transmission route of MERS coronavirus using decision tree and apriori algorithm. In 2016 18th International Conference on Advanced Communication Technology (ICACT), 2016.

[8] V. Lampas and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In 2010 2nd International Workshop on Cognitive Information Processing, pages 411–416, 2010.

[9] Huber Nieto-Chaupis. Face To Face with Next Flu Pandemic with a Wiener-Series-Based Machine Learning: Fast Decisions to Tackle Rapid Spread. In 2019 IEEE 9th Annual Computing

and Communication Workshop and Conference (CCWC), pages 654–658, 2019.

[10] M. Szomszor, P. Kostkova, and C. S. Louis. Twitter informatics: Tracking and understanding public reaction during the 2009 swine flu pandemic. In 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 1, pages 320–323, 2011.

[11] Yang Hong and Richard O. Sinnott. A social media platform for infectious disease analytics. In the 18th International Conference of Computational Science and Its Applications - ICCSA, volume 10960 of Lecture Notes in Computer Science, pages 526–540. Springer, 2018.

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

[13] T. Tran and K. Lee. Understanding citizen reactions and ebola-related information propagation on social media. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 106–111, 2016.

[14] C. Jason Wang, Chun Y. Ng, and Robert H. Brook. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. JAMA, 323(14):1341–1342, 2020**.**

[15] Xiangao Jiang, Megan Coffee, Anasse Bari, Jun-zhang Wang, Xinyue Jiang, Jianping Huang, Jichan Shi, Jianyi Dai, Jing Cai, Tianxiao Zhang, Zhengxing Wu, Guiqing He, and Yitong Huang. Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. CMC-Computers Materials & Continua, 63(1):537–551, 2020.

[16] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T. Gao, W. Duan, K. K. Tsoi, and F. Wang. Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. IEEE Transactions on Computational Social Systems, 7(2):556–562, 2020.

[17] Atika Qazi, Javaria Qazi, Khulla Naseer, Muhammad Zeeshan, Glenn Hardaker, Jaafar Zubairu Maitama, and Khalid Haruna. Analyzing Situational Awareness through Public Opinion to Predict Adoption of Social Distancing Amid Pandemic COVID-19. Journal of medical virology, 92(7):849–855, 2020**.**

Table 1. Performance of four sentiment prediction models using tf-idf features.

ML-Alg. Metric	Naive Bayes	Logistic Reg.	Decision Tree	Random Forest
Accuracy	**0.76**	**0.76**	0.71	**0.76**
Precision	**0.76**	**0.76**	0.75	0.74
Recall	0.88	0.88	0.78	**0.9**
F1-score	0.81	**0.82**	0.77	**0.82**

Table 2. Performance of four sentiment prediction models using BOW features.

ML-Alg. Metric	Naive Bayes	Logistic Reg.	Decision Tree	Random Forest
Accuracy	0.74	0.75	0.7	**0.77**
Precision	0.75	0.75	0.75	**0.76**
Recall	0.86	0.87	0.76	**0.88**
F1-score	0.8	0.8	0.75	**0.82**

Table 3. Performance of four sentiment prediction models using tf-idf features and stemming.

ML-Alg. Metric	Naive Bayes	Logistic Reg.	Decision Tree	Random Forest
Accuracy	**0.76**	0.75	0.7	**0.76**
Precision	**0.74**	0.73	0.72	0.72
Recall	0.88	0.89	0.78	**0.92**
F1-score	0.8	0.8	0.75	**0.82**

Table 4. Performance of four sentiment prediction models using BOW features and stemming.

ML-Alg. Metric	Naive Bayes	Logistic Reg.	Decision Tree	Random Forest
Accuracy	**0.78**	0.75	0.7	0.75
Precision	**0.78**	0.76	0.76	0.76
Recall	**0.87**	0.85	0.73	0.85
F1-score	**0.82**	0.8	0.74	0.8