International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 3
ISSN: 2088-5334

Machine Learning Model for Sentiment Analysis of COVID-19 Tweets

Malak Aljabri [a,b,*], Sumayh S. Aljameel [b], Irfan Ullah Khan [b], Nida Aslam [b], Sara Mhd. Bachar Chrouf [b], Norah Alzahrani [b]

[a] *Computer Science Department, College of Computers and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia*
[b] *Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia*

*Corresponding author: *msaljabri@iau.edu.sa*

***Abstract*— COVID-19 pandemic presents unprecedented challenges and enormously affects different aspects of individuals' lives worldwide. The implementation of different prevention measures, the economic and social disruption, and the significant rise in the mortality rate greatly affect the peoples' spectrum of emotions. Sentiment analysis, an important branch of artificial intelligence, uses machine learning techniques to understand public perspectives and gain more insights into how they think and feel. During the pandemic, sentiment analysis increasingly contributes towards making appropriate decisions. This research aims to analyze the public sentiment related to COVID-19 by exploring social perceptions shared on Twitter, one of the most ubiquitous social networks. This goal was achieved by building a machine learning model using a dataset of COVID-19 related English tweets. Different combinations of machine learning classification algorithms (Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB)) and feature extraction techniques (Term Frequency-Inverse Document Frequency (TF-IDF) and N-gram) were built and applied to the dataset for binary (positive, negative) and ternary (positive, negative, and neutral) classifications. A comparative study for the performance of the different models was then conducted, and the results concluded that XGB classification algorithm with unigram and bigram for binary classification achieved the highest accuracy of 90%. This sentiment analysis model can assist countries and governments in measuring the impact of the pandemic and the applied prevention measures on people's emotional and mental health and take early actions to reduce their impact or prevent them from becoming severe cases.**

***Keywords*— sentiment analysis; Twitter; COVID-19; machine learning.**

Manuscript received 18 Mar. 2021; revised 28 Dec. 2021; accepted 15 Feb. 2022. Date of publication 30 Jun. 2022.

I. INTRODUCTION

Social media has become a huge part of daily life, and its influence on public opinion is significant and reflects peoples' feelings, views, attitudes, and emotions about current issues and the latest news. COVID-19 pandemic is currently a devasting event affecting different aspects of individuals' lives worldwide. According to the latest statistics, the number of cases has reached approximately 100 million and 2.3 million deaths globally [1],[2]. Due to this pandemic, many people have experienced anxiety and fear of potential infection and infecting their families. In addition, social distancing, prevention measures, and other precautions taken by national governments and law enforcement agencies have caused depression and loneliness among many. Another notable effect of COVID-19 is the increased unemployment rates, which caused workers to worry about losing their jobs and livelihood opportunities [3]. As a result, the number of suicide cases has increased, especially in the eastern world [4]. Sentiment analysis has helped researchers and societies to understand public perspectives and gain deep insights into how people think and feel, especially with the huge number of sentimental posts being shared on social media every day. It also helps to develop and target marketing, by analyzing customer opinions and the latest trends [5]. Machine Learning (ML) and Natural Language Processing (NLP) techniques has led to a revolution in building models that can predict future events, classify data, and automated sentiment analysis [6].

During the COVID-19 pandemic, there is an exponential increase in the number of users, posts, and interactions among the people via social media. As it was the only platform where people could express their feelings, opinions, and communicate with each other [7],[8]. Community needs and

the sheer amount of data available on people's opinions have motivated us to build a model that analyzes the sentiments of Twitter posts, being one of the most popular social media platforms, rich with reviews and experiences related to COVID-19.

Analyzing Twitter data to examine public attitudes, concerns, and thoughts about the COVID-19 pandemic can take many research directions. One direction combines different topic modeling, NLP, ML, and sentiment analysis technologies to utilize the collected tweets to identify the key topics and themes within these trends and the sentiment associated with each topic. Various research has been conducted related to this direction that differs mainly regarding the goal, period of collected tweets, and the size of the dataset.

For example, a study conducted by Abraham et al. [9] examined the public response to mask-wearing prevention measures during the COVID-19 pandemic. They analyzed over 1 million COVID-19 related tweets that were specific to mask-wearing, collected over the first five months of 2020. All the tweets were clustered and divided into 15 high-level and 15 specific topic themes; each cluster was then subdivided into different categories based on the sentiment analysis. Then, an abstractive text summarization model was applied to provide summaries by combining related tweets for each sub-cluster. Moreover, the divisiveness of each tweet was computed weekly and globally, and, finally, linear regression of the divisiveness with time was carried out. The study discovered a noticeable increment in the volume of mask-related tweets between Mar. 17, 2020, and Jul. 27, 2020, proved that this increment was growing into negative sentimentality and polarity.

Chandrasekaran et al. [10] conducted a study to identify key trends in a dataset of over 13 million COVID-19 related tweets with a slightly different goal and dataset size. Tweets were collected between Jan. 1 and May 9, 2020, representing a combination of publicly available datasets and new datasets collected by the researchers. They identified the key trends in the overall dataset using the Latent Dirichlet Allocation (LDA) model and conducted a sentiment analysis to calculate weekly sentiment scores for 17 weeks using Valence Aware Dictionary and Sentiment Reasoner (VADER) techniques. The study revealed 26 topics classified into 10 themes. Among the analyzed tweets, 20.51% discussed the economic impact of COVID-19, 15.45% were on the spread of the virus, 13.14% were related to the treatment of the virus, 11.40% mentioned the effect on the healthcare industry, and finally, 11.19% focused on the government actions. The sentiment scores varied from positive to negative depending on the topics. For example, the topics spread, and growth of cases, symptoms, racism, and source of the outbreak had a negative average score, while those discussing prevention and impact on the economy and markets scored positively in the sentiment analysis process.

Xiang et al. [11] followed the same direction to identify the situations of older adults during the COVID-19 pandemic. They collected over 82,000 tweets between Feb. 1 and May 20, 2020. They applied supervised ML, topic modeling, sentiment analysis, and conventional statistical techniques. The frequencies of specific phrases related to older adults were counted, and then supervised ML was applied to the dataset to classify four distinct categories. Furthermore, LDA model was applied to extract 14 themes within the categories. Finally, Lexicon-based analysis was used to determine the sentiment of each tweet. The study indicated that 66.2% of tweets expressed personal opinions, 24.7% were informative, 4.8% contained jokes, and 4.3% reported personal experiences.

Using the same analysis approach and technologies, Boon-Itt and Skunkan [12] sought to reveal themes in English tweets and gain insight into public awareness of the COVID-19 pandemic. Over 107,000 tweets were collected and analyzed between Dec. 13 and Mar. 9, 2020. The study illustrated the trends in tweets during the pandemic. Moreover, the study presented an indication that Twitter users had a negative view of COVID-19. Finally, COVID-19 related themes were divided into three categories based on topic modeling techniques: the COVID-19 pandemic emergency, how to control COVID-19, and reports on COVID-19.

Xue et al. [13] identified psychological reactions to COVID-19 related topics among Twitter users and discourse between them. The researchers collected over 1.9 million COVID-19 related tweets from Jan. 23 to Mar. 7, 2020. They followed the same steps and used the same technologies as the two aforementioned studies but differed in applying unsupervised ML techniques. The study identified 11 latent topics related to COVID-19 and classified these into ten themes such as, "COVID-19 related death" and "cases outside China (worldwide)". The study also found that, due to the dynamic nature of COVID-19, fear of the virus was prevalent among all topics. In another study [14], the same researchers then collected a new dataset of over 4 million COVID-19 related tweets between Mar. 7 and Apr. 21, 2020, to identify unigrams, bigrams, salient topics, themes, and sentiments in the dataset. They identified 13 different topics and classified those topics into five themes: public health measures to slow the spread of COVID-19; social stigma associated with COVID-19; COVID-19 news, cases, and deaths; COVID-19 in the United States; and COVID-19 in the rest of the world. Unlike all other topics, the study revealed that people demonstrated fear when discussing new cases or deaths due to COVID-19. Moreover, among the other topics, the prevailing sentiment was the anticipation of measures being announced to address COVID-19. The study also identified popular unigrams such as, "virus" and "lockdown", and popular bigrams such as "COVID-19" and "stay home".

A study conducted by Medford et al. [15] aimed to comprehend the changes in sentiment, activities, and content on Twitter during the COVID-19 pandemic. They created a dataset comprising 126,049 tweets and used a word cloud to identify the most common words related to COVID-19. They also used the Syuzhet R package for sentiment analysis and the Recurrent Neural Network (RNN) method to label the emotions based on Ekman's emotional categories and LDA for topic modeling. Their results showed that 49.5% and 30% of tweets demonstrated fear and surprise emotions. Furthermore, political and economic topics were the most commonly discussed, while public health risks and countermeasures were the least discussed topics.

Another research direction classifies the sentiment of tweets into positive and negative by building learning models

and classifiers using ML or deep learning techniques. With a similar goal to our research, Nemes and Kiss [16] conducted a study on emotion predictions regarding COVID-19. They collected and analyzed 500 tweets posted between Apr. 24 and May 13, 2020; then, NLP and sentiment classification RNN techniques were used to identify users' tweets' emotions, establish connections between words, and label them into different classes. To overcome the limitation of the small size of the dataset, they added further emotional classes to their classifications and the two extreme classes of positive and negative. These classes were: weakly positive/negative, and strongly positive/negative emotions. The study proposed that, despite the negativity related to COVID-19 that was expressed on social media, the overall positivity did not disappear and was still present during the pandemic.

Along with the growth in the number of tweets posted in relation to COVID-19 and the World Health Organization (WHO), there has been an increasing number of misleading tweets, representing a source of psychological and emotional stress. Chakraborty et al. [17] studied the effects of the rapid spread of tweets containing false information tweets. They created two datasets and classified the tweets into positive, negative and neutral. The first dataset contained 23,000 tweets consisting of mostly re-tweeted tweets from January 2020 to March 2020, with the majority of tweets being neutral or negative. The second dataset contained 226,668 tweets from December 2019 to May 2020, with most tweets being positive or neutral. The researchers applied a fuzzy rule-based model with a Gaussian membership function. After collecting and pre-processing the tweets, they implemented Bag Of Words (BoW) as the feature extraction technique, with different settings such as count vectorizer, unigram, bigram, trigram, Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer, and Doc-2-Vec. For tweet labelling, they used the TextBlob and Afinn modules. The model achieved an accuracy of 81% with the first dataset using Logistic Regression (LR) with trigrams under the TF-IDF vectorizer. In comparison, the second dataset achieved an accuracy of 75% also using LR but with bigrams under the TF-IDF vectorizer. The study concluded that people tend to post positive tweets and re-tweet negative tweets.

Using the Arabic twitter dataset, Aljabri et al. [18] examined the public opinions regarding the distance learning applied in Saudi Arabia during the COVID-19 pandemic. The dataset used consisted of 20,827 tweets divided into different education sectors. The pre-processed the dataset and applied a combination of TF-IDF for feature engineering. Then conducted a comparative performance study and concluded that LR with unigram under the TF-IDF vectorizer achieved the best performance accuracy of 89.9%.

In another study, Samuel et al. [19] analyzed the public sentiment toward COVID-19 using ML techniques and studied the effect of tweet-length on the classification accuracy. Over 900,000 tweets were collected from February to March of 2020. The researchers classified the tweets positively and negatively and used N-gram for vectorization. Specifically, they applied unigram, bigram, trigram, and quad-grams sequences. They also analyzed the geographical location of the tweets posted and the user location. For classification, they used Naïve Bayes (NB) and LR. The proposed model achieved an accuracy of 91% using NB and an accuracy of 74% using LR with short tweets. However, both models showed a weak performance with long tweets.

Kruspe et al. [20] developed a method that involved analyzing tweets collected during the first months of the COVID-19 pandemic in Europe. The researchers aimed to automate the tweet's sentiment to analyze the changes in sentiment over time based on country. To build the model, they used Neural Network (NN) architecture and the Multilingual Universal Sentence Encoder (MUSE) for sentence-level embedding and trained the model with the Sentiment140 dataset that consists of 1.5 million tweets. They then used a dataset of 4.6 million multilingual tweets collected and found that the sentiment of the tweets started out negative and then became more positive toward the end of the dataset period. Apart from Germany, the development of sentiment remained significantly below the average in all countries.

Unlike the work mentioned above, this research aims to conduct a sentiment analysis study to analyze the community sentiment related to COVID-19 by exploring social perceptions shared on Twitter. This goal was achieved by building a machine learning (ML) based model using a dataset of English tweets. Three ML classification algorithms were applied to the dataset: Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB), in order to compare and analyze their performance. TF-IDF and N-gram were used as feature extraction and selection techniques.

The main contributions of this research are:

- Build different sentiment analysis models on a dataset of COVID-19 related tweets. The models consist of different combinations of ML classifiers (XGB, SVM, and RF), and features extraction techniques (Unigram, Bigram, Unigram with TF-IDF, and Bigram with TF-IDF), and different classifications classes (positive-negative, or positive-negative – neutral).
- Conduct comparative study on the performance of the different model on COVID-19 dataset and concluded that XGB classifier with Unigram and binary classification achieved the highest accuracy of 90%.

The remainder of this paper is divided into the following sections: Section 2 presents our research methodology and the related stages; Section 3 explains the performance evaluation; Section 4 concludes and summarizes the main contributions of this paper and discusses future works.

II. Materials and Method

Fig. 1 presents the steps of the proposed framework to build sentiment analysis model. As indicated in the figure, after downloading the dataset, we commenced the pre-processing step, including various data cleaning, text normalization, and standardization techniques. Following the pre-processing, the features were selected and extracted using a combination of N-gram and TF-IDF approaches to produce vectors. Then, we applied a set of ML classification algorithms to classify the tweets as positive, negative, or neutral. Finally, the evaluation and performance comparison processes were carried out in terms of recall, precision, F-score, and accuracy. These steps are described in detail in the following subsections.

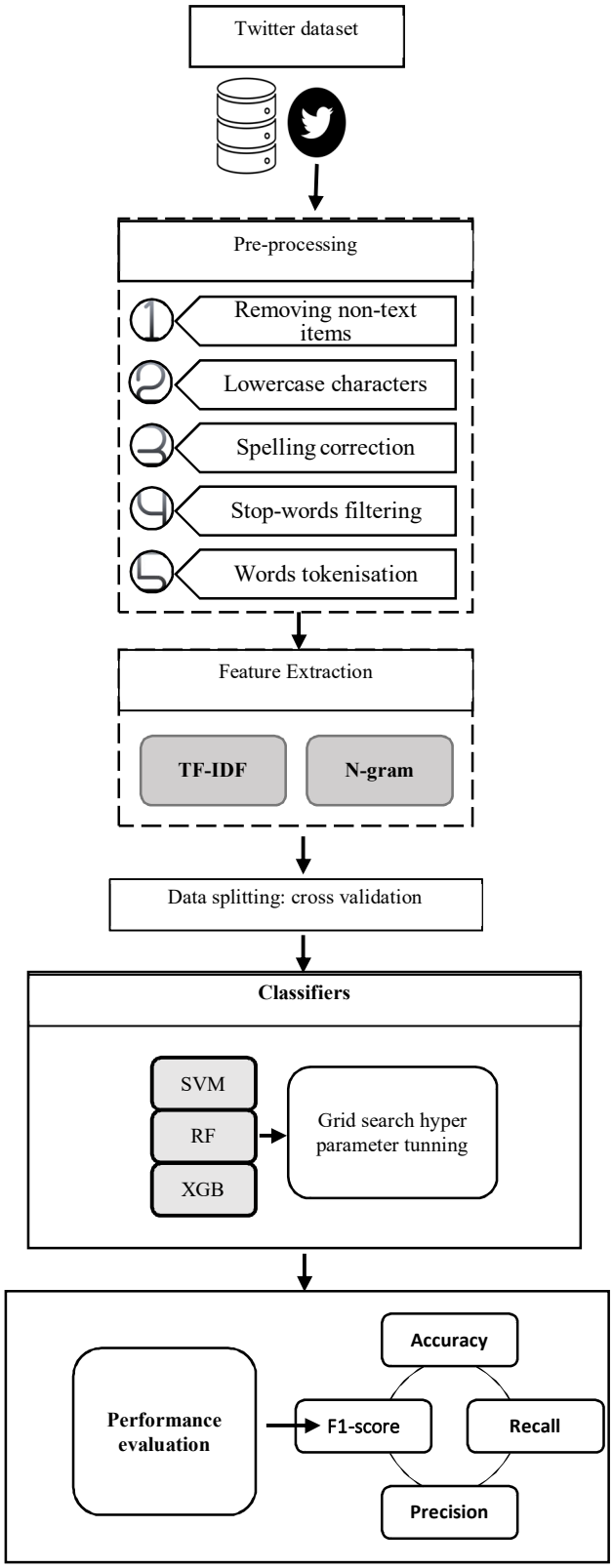

Fig. 1 Research Methodology

A. Dataset Description and Exploration

In our study, we built sentiment analysis models and conducted a comparative study using the Kaggle dataset.[21] This dataset contains 44,478 English tweets crawled from Dec. 31, 2019, to Sept. 8 2020. The tweets were manually classified and labelled according to three sentiment classes (positive, neutral, negative), as follows: 16,976 neutral sentiments; 19,505 positive sentiments; and 7,997 negative sentiments. Table 1 shows a sample of the positive, negative, and neutral tweets.

In our sentiment analysis models performance comparison, we investigated the performance of the models in two setups:

- We studied the performance considering two classification classes (positive and negative); in this setup, a set of 27,000 tweets was used.
- We studied the performance considering three classification classes (positive, negative, and neutral); in this setup, the complete set of 44,000 tweets was used.

TABLE I
EXAMPLE OF LABELED TWEETS

Tweet	category
Hi, COVID-19. Thanks for making me do more online shopping.	Positive
Pausing student loan payments and halting interest accumulation amp stopping punitive student loan collections would provide much needed immediate relief to those unable to work amp are facing economic hardship.	Negative
Please don't hoard food and water. There's absolutely no need to panic buy; the supply chain is completely interrupted. And above all, please don't hoard sanitizing products; there are people out there who really need them, probably more than you. #DontPanicBuy #coronavirus	Neutral

We conducted an exploratory analysis of the dataset as follows:

1) Keyword Trend Analysis: We performed keyword trends study on the dataset and found that the most common word in our analysis was "COVID" which appeared in the dataset 11329 times. The prevalence of this word in the dataset is because all the tweets were related to the COVID-19. Table 2 shows the 20 most common words in the dataset and the frequency of their occurrences, which indicated that people talked about prices, food, shopping, and online as their main concerns during the pandemic.

TABLE II
TOP-20 FREQUENT WORDS FROM THE DATASET

Word	Count
COVID	11329
prices	8285
food	8209
store	7908
supermarket	7221
grocery	6909
people	6401
consumer	4628
shopping	3880
online	3864
get	3202
need	3077
panic	3063
pandemic	2822
stock	2775
us	2552
go	2543
home	2591

Word cloud is a library used to visualize and display the most frequently used words in the dataset. Fig. 2 presents the most frequent words in the dataset. Fig. 3 and Fig. 4 present the most frequent words in the dataset with positive and negative labels, respectively.

Fig. 2 Word cloud of common words in the dataset.

Fig. 3 Word cloud of positive tweets.

Fig. 4 Word cloud of negative tweets.

2) Topic Modeling: For quantitative analysis of topics in the dataset, we applied the LDA model which generated topics from the tweets. We set the number of topics to 10 and presented the top 20 words that contributed to each topic group, as demonstrated in Table 3.

TABLE III
WORDS CONTRIBUTED TO EACH TOPIC

Topic	Words Contributed to Topic
1	workers, work, retail, staff, working, stores, health, essential, customers, employees, store, risk, open, care, hours, public, delivery, service, medical, closed
2	COVID, time, times, free, support, crisis, great, companies, businesses, small, real, leave, share, coming, retailers, sick, community, normal, company, response
3	prices, world, market, amid, pandemic, global, outbreak, economy, year, markets, economic, trump, China, drop, production, march, month, years, fall, rise
4	store, grocery, toilet, paper, today, line, life, long, days, stores, roll, lines, left, morning, yesterday, stuff, kids, wait, bread, crazy
5	people, supermarket, panic, buying, shelves, local, stop, shop, things, supermarkets, shops, elderly, hoarding, items, essentials, thing, selfish, fresh, weeks, stocked
6	food, stock, demand, supply, COVID, supplies, lockdown, money, chain, country, increased, vulnerable, banks, bank, items, week, families, running, shortage, meet
7	COVID, coronavirus, price, prices, products, government, high, news, check, taking, state, outbreak, situation, selling, scams, information, report, protection, protect, related
8	shopping, online, home, sanitizer, hand, stay, safe, delivery, order, making, hands, family, weeks, week, water, fight, groceries, avoid, house, wash
9	consumer, COVID, pandemic, crisis, impact, read, business, consumers, industry, post, latest, behavior, data, change, spending, financial, learn, April, update, Americans
10	supermarket, virus, social, good, masks, spread, people, face, mask, distancing, gloves, place, live, cases, today, person, shoppers, daily, corona, measures

The word cloud of words within top ten topics is demonstrated in Fig. 5.

Fig. 5 Word cloud of the top 10 topics.

B. Data Cleaning and Pre-processing

The most important step in a sentiment analysis study is preparing the tweets for analysis. Raw tweets are not in a suitable format for analysis as they often are written in informal language, contain many non-text symbols, lots of noise, and irrelevant text such as advertisements, emoticons, folksonomies, and slang. Additionally, many of the words in such tweets have no effect on their overall sentiment classification.

The cleaning and pre-processing of tweets are crucial for the creation of sentiment analysis and directly impact the

model's accuracy. In this research, we applied the following pre-processing steps to the dataset:

1) Removing non-text items: including all hashtags, mentions, symbols, emojis, links, and pictures.

2) Lowercase characters: converting all characters in tweets to lowercase.

3) Spelling correction: applied to all tweets using the JamSpell [22] library. Steps 2 and 3 play an important role in optimizing the models' efficiency by avoiding the same words being recognized as different due to spelling mistakes or character capitalization differences.

4) Stop-words filtering: removing all stop-words, which are words that do not contribute to the overall meaning and sentiment of the tweet ('me', 'my', 'who', 'the').

5) Words tokenization: each tweet was broken into smaller parts, called tokens.

C. Feature extraction

Sentiment analysis requires a model that classifies the emotions and opinions expressed in the text. However, ML classifiers can only handle numerical data; therefore, there is a need to extract the text features into numerical vectors. Two of the most applied techniques to extract textual features are N-gram and TF-IDF.

The N-gram technique represents the text as an N-words sequence; it can be simple or complex, based on the value of N. In unigrams, it considers each word a sequence, while in bigrams it considers each pair of words a sequence. Then, the vectorizer calculates the occurrences of each sequence to generate the sentences' vectors [23].

Another commonly used technique is TF-IDF, which extracts features by giving weights. It focuses on two main values:

- The first value is the term frequency (TF), which represents a local weighting scheme by calculating each term's occurrence in a document (or tweet).
- The second value is the inverse document frequency (IDF), which represents a global weighting scheme. It gives the logarithmic value of the total number of documents (tweets) divided by the number of documents (tweets) a term has appeared in.

Equations 1 and 2 represent the formulas used to calculate each of these two values:

$$\mathrm{TF(t)} = \frac{\text{number of the term's occurance}}{\text{total number of the terms in the document (tweet)}} \quad (1)$$

$$\mathrm{IDF(t)} = \log\left(\frac{\text{number of documents(tweets)}}{\text{number of documents (tweets)that contain the term}}\right. \quad (2)$$

After computing the TF and IDF for each term, the TF-IDF (term weight) is calculated by multiplying the TF by the IDF values. This will produce a lower weight if the term frequently appears in every tweet in the set and a greater weight for uncommon terms [24].

These techniques are widely used to build classification models; in our study, we built different ML models using different feature extraction techniques, as below:

- N-gram technique with the value of N equal to 1 and 2.
- A combination of TF-IDF and N-gram techniques.

D. Data Splitting

ML models are built using a training set, where the goal is to learn the pattern of the data to generalize the model to new and unknown data. The second set is the testing set, where unseen data are fed into the model to predict the output and analyze that predicted output to evaluate the model's performance.

This study divided the dataset into a training set and testing set using the K-fold cross-validation method. This method splits the data into K-folds and, in each iteration, one fold will be the testing set and the other folds will be the training set. After evaluating the performance measure values for each iteration, the average of the values will be calculated to evaluate the model. This experiment used three folds in the grid search and the final model evaluation.

E. Classification Algorithm

We applied a set of ML algorithms for sentiment classifications to conduct a comprehensive analysis, tuned their performance, and compared the results. The set of classifiers employed in our study was: XGBoost (XGB), SVM, and RF. In this section, we discuss a brief description and the optimized parameters of each classifier.

1) XGB Classification Algorithm: The XGB algorithm, extreme gradient boosting, is an ML classifier based on the gradient tree-boosting technique. It is known for fast learning and performance scalability, as it uses parallelism and enables different algorithm optimizations. The 'boosting' element refers to the fact that this algorithm uses the ensemble method to enhance the model's performance further. The 'gradient' refers to the fact that it uses previous errors to enhance future models. XGB is widely used and has achieved excellent results in challenging problems in ML [25], [26], [27]. We applied XGB for sentiment classification and conducted extensive experiments to optimize its performance by tuning its parameters using grid search to produce hyper-optimized parameters.

2) SVM Classification Algorithm: The SVM classifier is a supervised ML algorithm that is used to solve many different problems, such as classification, regression, detection, and feature selection. One of the main aspects of SVM is calculating the best hyperplane that maximizes the separation of data to each class. When dealing with linear data, the separation is easy; however, when the data becomes more complex and non-linear, the kernel functions are used to map the data to feature space, where the data can be linearly separated. Other parameters create the boundaries for the hyperplanes in the SVM [28],[29]. We applied SVM for sentiment classification and conducted extensive experiments to optimize its performance by tuning the parameters using grid search.

RF Classification Algorithm: The RF classifier is a supervised classifier, which, as the term implies, uses a forest of trees. It is an ensemble algorithm that creates number of decision trees, where each tree has a sample of data. Then, these trees are voting to choose the best classification tree. The RF algorithm is considered to be un-biased, since each tree works on a sample of data, and the method works well with missing data[30],[31]. We applied RF for sentiment

classification and conducted extensive experiments to optimize its performance using grid search optimization. Table 4 shows the parameters used with XGB, SVM, and RF.

TABLE IV
PARAMETERS USING GRID SEARCH OPTIMIZATION

Classifier	Parameter	Assigned Value
XGBoost	Objective	binary: logistic
	learning_rate	0.1
	max_depth	10
	min_child_weight	1
	Subsample	0.7
	colsample_bytree	0.7
	n_estimators	500
SVM	C	5
	Kernal	'linear'
	degree	3
	coef0	1
	gamma	5
RF	n_estimators	800
	max_depth	100
	min_samples_split	2
	min_samples_leaf	1
	n_estimators	1

F. Performance Measurements

To evaluate the performance of our models, four different performance measures were used [32],[33]. First, the accuracy was measured, as the number of correctly predicted sentiments divided by the total number of all the predicted sentiments. Accuracy was calculated using Equation 3.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (3)$$

Where:

- True Positives (TP) refers to the number of tweets that are predicted as positive and are correctly positive,
- True Negatives (TN) refers to the number of tweets predicted as negative and correctly negative.
- False Positives (FP) refer to the tweets predicted as positive but correctly negative.
- False Negatives (FN) refer to the tweets predicted as negative but correctly positive.

Second, the precision was measured by calculating the false positives of the classifier. Precision was calculated using Equation 4.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

Third, recall was measured by calculating the false negatives of the classifier. Recall was calculated using Equation 5.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (5)$$

Finally, the F1-score was calculated by taking the weighted harmonic average of the recall and precision. F1-score was calculated using Equation 6.

$$\text{F1} - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

III. RESULTS AND DISCUSSION

In this section, we measure the performance and present the results of the sentiment classification models discussed in the sections above. We compare the performance of the models that represent combinations of different sentiment classes (positive and negative; or positive, negative, and neutral), four text feature extraction techniques (unigram, bigram, unigram with TF-IDF, and bigram with TF-IDF), and three tuned classifiers (XGB, SVM, RF). We conducted our experiments in two main stages, as follows:

1) Classification into three classes (positive, negative, and neutral): In the first stage, we studied the performance of combinations of the three aforementioned classifiers and four feature extraction techniques to classify the data into three classes. Table 5 illustrates the findings extracted. The XGB classifier outperformed the SVM and RF classifiers, as it achieved the highest values: an accuracy of 82%, a precision of 83%, a recall of 82 %, and an F1-score of 82 % using the bigram technique. Additionally, the bigram technique achieved the best results for the SVM classifier, but its performance decreased when applied with RF.

TABLE V
MODELS' RESULTS USING THREE CLASSES

Classifier	Feature Extraction	Acc	Prec	Rec	F1-score
SVM	unigram	78	78	78	78
	bigram	80	80	80	80
	unigram + TF-IDF	80	80	80	80
	bigram + TF-IDF	78	78	78	78
RF	unigram	66	72	66	60
	bigram	63	46	51	46
	unigram + TF-IDF	69	71	69	66
	bigram + TF-IDF	63	72	63	57
XGB	unigram	82	82	82	82
	bigram	82	83	82	82
	unigram + TF-IDF	80	80	80	80
	bigram + TF-IDF	80	80	80	80

2) Classification into two classes (positive and negative): In the second stage, we studied the performance of combinations of the three classifiers above and four feature extraction techniques to classify the data into two classes. Table 6 illustrates the findings extracted. In general, the N-gram results were better without using the TF-IDF technique for all classifiers. The best results were mostly achieved using the XGB classifier, with Unigram/bigram feature extractions methods with the accuracy of 90%, and 94%, 91%, and 93% in precision, recall, and F1-score, respectively. On the other hand, the RF classifier performed poorly with all feature extraction techniques but had the highest recall values, reaching 100% using bigram; this indicates that this model does not give any false negative predictions, even though it produced the worst values with other performance measures.

To conclude, we can state that XGB for sentiment analysis outperforms SVM and RF classifiers. In addition, when using binary classification, all three classifiers showed a significantly improved performance. In particular, the XGB classifier showed an improved accuracy level, from 82 % to 90%. In terms of feature extraction techniques, in all

conducted experiments, TF-IDF did not perform well compared with N-gram.

In summary, the main result achieved by our experiments was an accuracy of 90% using XGB with unigram or bigram on binary classification, applied to a new COVID-19 dataset, to build sentiment analysis models. It is worth mentioning that this dataset [21]was used by Silva et al. [34] to predict misleading information about COVID-19, and an average F1-score of 82% was found. By contrast, our study aimed to predict sentiment, and we achieved a F1 score of 93%.

TABLE VI
MODELS' RESULTS USING TWO CLASSES

Classifier	Feature Extraction	Acc	Prec	Rec	F1-score
SVM	unigram	87	92	89	91
	bigram	88	93	89	91
	unigram + TF-IDF	88	92	91	91
	bigram + TF-IDF	86	88	93	90
RF	unigram	78	77	99	86
	bigram	71	71	100	83
	unigram + TF-IDF	78	77	98	86
	bigram + TF-IDF	72	72	100	83
XGB	unigram	90	94	91	93
	bigram	90	94	91	93
	unigram + TF-IDF	88	92	92	92
	bigram + TF-IDF	88	92	91	92

IV. CONCLUSION

This study aimed to build and compare the performance of several ML models for the analysis of Twitter users' sentiments regarding COVID-19. The experiments were conducted by building different models consisting of different combinations of classifiers, feature extraction techniques, and number of class labels. We studied the performance of SVM, RF, and XGBoost with N-gram and TF-IDF for feature extraction. We also analyzed the effect of using binary classification (positive or negative) and classification to three classes (positive, negative, or neutral). The study showed that XGBoost outperforms SVM and RF in analyzing sentiments, as it achieved 90% accuracy in the binary class dataset with unigrams and bigram. In addition, we can say that binary classification performed better than multi-class classification.

The sentiment analysis model can help countries and governments measure the pandemic's effect and precautionary measures on citizens' emotional and mental health and take early action to aid citizens in improving their mental health. Moreover, mental health care services and the health sector can apply the model to measure the mental status of the society and take rapid action to reduce mental and emotional issues and prevent them from becoming severe cases that need medical intervention, for example suicidal actions.

In the future, we plan to apply our model to new COVID-19 related tweets to study sentiment changes with any changes related to the spread of COVID-19 or the implementation of new prevention measures. Moreover, we may explore different promising research directions for sentiment analysis. In particular, we plan to study the performance of deep learning (DL) classifiers and compare their performance with ML classifiers and apply different feature extraction techniques.

ACKNOWLEDGMENT

The Deanship of Scientific Research supported this work, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia [Grant No. COVID19-2020-060-CSIT].

REFERENCES

[1] "COVID-19 Coronavirus Pandemic". Accessed on Mar. 02, 2021, [Online]. Available: https://www.worldometers.info/coronavirus/.

[2] "Cumulative Cases". Accessed on Mar. 02, 2021, [Online]. Available: https://coronavirus.jhu.edu/data/cumulative-cases.

[3] B. Semo and S. M. Frissa, "The Mental Health Impact of the COVID-19 Pandemic: Implications for Sub-Saharan Africa," *Psychology Research and Behavior Management*, vol. 13, pp. 713–720, Sep. 2020, Accessed on Mar. 02, 2021, DOI: 10.2147/PRBM.S264286.

[4] T. Tanaka and S. Okamoto, "Increase in suicide following an initial decline during the COVID-19 pandemic in Japan," *Nature Human Behaviour*, no. 5, pp. 229–238, 2021, Accessed on Mar. 02, 2021, DOI:10.1038/s41562-020-01042-z.

[5] R. Wagh and P. Punde, "Survey on Sentiment Analysis using Twitter Dataset," Presented at *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018*, pp. 208–211. DOI: 10.1109/ICECA.2018.8474783.

[6] T. Beysolow II," What Is Natural Language Processing?" in *Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing* ,1st ed. New York, NY, USA, Apress 2018, pp.1-12, [Online]. Available: https://doi.org/10.1007/978-1-4842-3733-5_1

[7] S. Dubey et al., "Psychosocial impact of COVID-19," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 14, no. 5, pp. 779–788, 2020, Accessed on Mar. 02, 2021, DOI: 10.1016/j.dsx.2020.05.035.

[8] M. Cinelli et al., "The COVID-19 social media infodemic," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020, Accessed on Mar. 02, 2021, DOI:10.1038/s41598-020-73510-5.

[9] A. C. Sanders et al., "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse," *medRxiv*, 2020, Accessed on Mar. 02, 2021, DOI:10.1101/2020.08.28.20183863.

[10] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, "Topics, Trends, and Sentiments of Tweets about the COVID-19 Pandemic: Temporal Infoveillance Study," *Journal of Medical Internet Research*, vol. 22, no. 10, pp. 1–12, 2020, Accessed on Mar. 02, 2021, DOI: 10.2196/22624.

[11] X. Xiang et al., "Modern Senicide in the Face of a Pandemic: An Examination of Public Discourse and Sentiment About Older Adults and COVID-19 Using Machine Learning," *The Journals of Gerontology: Series B*, vol. XX, no. Xx, pp. 1–11, 2020, Accessed on Mar. 02, 2021, DOI: 10.1093/geronb/gbaa128.

[12] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study," *JMIR Public Health and Surveillance*, vol. 6, no. 4, pp. 1–17, 2020, Accessed on Mar. 02, 2021, DOI:10.2196/21978.

[13] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter," *Plos one*, vol. 15, no. Sept. 9, pp. 1–12, 2020, Accessed on Mar. 02, 2021, DOI: 10.1371/journal.pone.0239441.

[14] J. Xue et al., "Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach," *Journal of Medical Internet Research*, vol. 22, no. 11, pp. 1–14, 2020, Accessed on Mar. 02, 2021, DOI:10.2196/20550.

[15] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, "An 'Infodemic': Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak," *Open Forum Infectious Diseases*, vol. 7, no. 7, 2020, Accessed on Mar. 02, 2021, DOI: 10.1093/ofid/ofaa258

[16] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *Journal of Information and Telecommunication*, pp. 1–15, 2020, Accessed on Mar. 02, 2021, DOI:10.1080/24751839.2020.1790793.

[17] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Applied Soft Computing Journal*, vol. 97, 2020, Accessed on Mar. 02, 2021, DOI: 10.1016/j.asoc.2020.106754.

[18] M. Aljabri et al., "Sentiment analysis of arabic tweets regarding distance learning in saudi arabia during the COVID-19 pandemic," *Sensors*, vol. 21, no. 16, 2021, Accessed on Mar. 02, 2021, DOI: 10.3390/s21165431.

[19] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," *Information (Switzerland)*, vol. 11, no. 6, pp. 1–22, 2020, Accessed on Mar. 02, 2021, DOI: 10.3390/info11060314.

[20] H. Matthias, A. Kruspe, and I. Kuhn, "Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic", arXiv preprint arXiv:2008, July.21,2020,[Online]Available: https://arxiv.org/pdf/2008.12172.pdf.

[21] A. Miglani, "Coronavirus tweets NLP - Text Classification | Kaggle." Accessed on Mar. 02, 2021, [Online] Available: https://www.kaggle.com/datatattle/COVID-19-nlp-text-classification?select=Corona_NLP_test.csv.

[22] "JamSpell". Accessed on Mar. 02, 2021, [Online] Available: https://github.com/bakwc/JamSpell. Accessed on Mar. 02, 2021.

[23] S. S. M. M. Rahman, K. B. M. B. Biplob, M. H. Rahman, K. Sarker, and T. Islam, "An investigation and evaluation of N-gram, TF-IDF and ensemble methods in sentiment classification," in *In: Bhuiyan T., Rahman M.M., Ali M.A. (eds) Cyber Security and Computer Science. ICONCS 2020.* Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 325. Springer, Cham, pp. 391–402, 2020, [Online]. Available: DOI: 10.1007/978-3-030-52856-0_31.

[24] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018, Accessed on Mar. 02, 2021, DOI: 10.5120/ijca2018917395.

[25] Z. Li, Q. Zhang, Y. Wang, and S. Wang, "Social media rumor refuter feature analysis and crowd identification based on XG Boost and NLP," *Applied Sciences (Switzerland)*, vol. 10, no. 14, 2020, Accessed on Mar. 02, 2021, DOI: 10.3390/app10144711.

[26] Y. Jia et al., "GNSS-R soil moisture retrieval based on a XGboost machine learning aided method: Performance and validation," *Remote Sensing*, vol. 11, no. 14, pp. 1–25, 2019, Accessed on Mar. 02, 2021, DOI:10.3390/rs11141655.

[27] L. Zhang and C. Zhan.(2017, April) "Machine Learning in Rock Facies Classification: An Application of XGBoost,"presented at *International Geophysical Conference*. Accessed on Mar. 02, 2021, [Online]. Available: https://doi.org/10.1190/IGC2017-351.

[28] V. Jakkula, "Tutorial on Support Vector Machine (SVM)," School of EECS, Washington State University, pp. 1–13, 2011. Accessed on Mar. 02, 2021, [Online]. Available: https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf.

[29] M. Awad and R. Khanna, "Support Vector Machines for Classification" in *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, Apress, Berkeley, CA, 2015, pp. 67-80, [Online]. Available: https://doi.org/10.1007/978-1-4302-5990-9_4.

[30] Z. Xiong, X. Sun, J. Sang, and X. Wei, "Modify the Accuracy of MODIS PWV in China : A Performance Comparison Using Random Forest , Generalized Regression Neural Network and Back-Propagation Neural Network," *Remote Sensing*, vol. 13, no. 11, pp. 1–18, 2021, Accessed on Mar. 02, 2021, DOI: 10.3390/rs13112215.

[31] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, 2019, Accessed on Mar. 02, 2021, DOI: 10.1002/widm.1301.

[32] A. Burkov, "Model Performance Assessment" in *The Hundred-Page Machine Learning Book*, *Illustrate*. Andriy Burkov, 2019.

[33] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation,". In: *Losada D.E., Fernández-Luna J.M. (eds) Advances in Information Retrieval*, ECIR 2005. Lecture Notes in Computer Science, vol 3408. Springer, Berlin, Heidelberg, 2005, pp. 345-359, [Online]. Available: https://doi.org/10.1007/978-3-540-31865-1_25

[34] M. Silva et al., "Predicting misinformation and engagement in COVID-19 twitter discourse in the first months of the outbreak," arXiv, vol. 37, no. 4, 2020.