ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 12 (2022) No. 6, pp. 50-69
DOI:10.33168/JSMS.2022.0604

Sentiment Analysis of Covid-19 Tweets by Supervised Machine Learning Models

Aina Afrina Binti Mohd Nasir, Naveen Palanichamy[+]

Faculty of Computing and Informatics, Multimedia University (MMU), Cyberjaya, Malaysia

p.naveen@mmu.edu.my

Abstract. The COVID-19 virus's transmissibility has sparked intense debate on social media sites, particularly Twitter. As a result, to employ resources efficiently and effectively, a comprehensive assessment of the situation is crucial. Therefore, COVID-19 tweet sentiment analysis is implemented in this research by employing a supervised machine learning (ML) approach. Data is retrieved from Twitter using the Tweepy API, pre-processed using pre-processing techniques, and sentiment extracted and labelled as positive or negative sentiments using the TextBlob library. Three separate feature extraction techniques are used: Bag-of-words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) combination with 1-gram, and TF-IDF combination with 2-gram. The sentiment is then analyzed using ML classifiers such as Random Forest (RF) and Support Vector Machine (SVM). For clarity, the dataset is studied further using the deep learning method which is Long Short-Term Memory (LSTM) architecture. The four standard evaluation metrics, Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC) were used to evaluate the performance of the models. The findings show that the RF classifier surpasses all other models with a 0.98 accuracy score when combining 2-gram TF-IDF features. In summary, the model may be used to categorize perspectives and will assist policymakers in making more educated decisions about how to respond to the current pandemic.

Keywords: supervised machine learning, random forest, support vector machine, feature extraction

1. Introduction

COVID-19 had already arisen as a harmful and dangerous public health issue worldwide. People are dying all around the world because of the COVID-19 pandemic. This sickness has had profound effects on people in both explicit and tacit ways. The government has enacted new rules such as staying at home, social isolation, and restrictions on people's movement to combat the disease's spread. The internet is the primary source of communication with the rest of the world in this situation.

Individuals' lives are influenced by social media, which connects them to the rest of the world. It is impossible to function without access to social media to stay up to date on the latest news, such as coronavirus updates. People nowadays rely more on posts and tweets shared on social media sites such as Facebook, Instagram, and Twitter. For example, due to current health emergencies, all organizations, individuals, and governments used and rely on social media platforms to communicate with each other. As a result, traffic on social networking sites has skyrocketed. Furthermore, people are frequently using Twitter to exchange their opinions and obtain necessary information during this COVID-19 outbreak. Therefore, Twitter has outperformed its competitors in terms of disseminating Covid-19 information promptly.

It is expected that social media posts will direct individuals to receive accurate and reliable information. However, in many cases, such as the COVID-19 material that spread on social media, the information led to false conclusions. When you look at the coronavirus posts, it is evident that they misled individuals by providing inaccurate facts. In other recent epidemics, such as Ebola, misinformation was widely disseminated. This is a disturbing trend since many individuals mistakenly believe that disinformation is accurate information.

People's minds had already been upset by the coronavirus; now, comments and tweets about COVID- 19 are unsettling and a source of concern that needs to be addressed to deal with misleading information from many sources. As a result, all citizens are physically and psychologically impacted.

As a result, the fast-moving COVID-19 pandemic requires thorough investigation and identification of misinformation. At the same time, this scenario has piqued the interest of researchers, who are exploring using sentiment analysis to develop a more comprehensive picture. Therefore, the main goal of this study is to acquire a better understanding of public perspectives and opinions on COVID-19, as well as to refute misunderstandings about the epidemic through sentiment analysis using Twitter data. The main reasons for performing sentiment analysis on Twitter would be that Twitter is extremely useful for extracting information about a user's thoughts, ideas, and insights on a wide range of issues, as well as due to post length limitations.

In conclusion, tweets collected from Twitter data for sentiment analysis of persons affected by the coronavirus using machine learning algorithms will aid in classifying users' sentiments into two categories during the disease outbreak: positive and negative. In short, this study employs supervised machine learning methods to analyse sentiment on the acquired COVID-19 tweets dataset.

2. Related Work on Sentiment Analysis on Tweets

This section of the paper focuses on the literature review on the types of feature extraction, machine learning approaches, and performance metrics used to perform sentiment analysis by various researchers.

2.1. Feature Extraction

In the first paper by (Kastrati et al., 2021), BoW, specifically TF and TF- IDF, are employed as feature extraction representations in traditional ML models on sentiment classification projects. Each Twitter message was given a positive, negative, or neutral value in both studies. Following that, for the experiment by (Aljabri et al., 2021), different models were created by combining different N- gram sizes (unigram and bigram) with the TF-IDF technique to turn textual data into numerical variables that the algorithms could analyze and operate with. Also, (Amin et al., 2021) used the help of the TF-IDF feature extraction technique to perform sentiment classification on the collected tweets which are around 900,000 tweets.

In (Ali, 2021), the authors used Information Gain (IG) as a filtering strategy to increase classification performance and BoW for feature extraction to improve classification performance. Other than that, (Aribowo et al., 2020) employed two well-known feature extraction algorithms, CountVectorizer (CV) and TF- IDF. (Zhang et al., 2020) used two of the most common feature extraction approaches, N-gram, and TF-IDF on previously gathered twitter data. Different feature approaches by (Khan et al., 2021), such as BoW and TF-IDF, are employed in the experiment to preserve expressive information. Lastly, the authors of this paper (Sajib et al., 2019**)** focus on N-gram approaches utilizing unigrams and bigrams features.

As shown in Table 1, out of five feature extraction, TF-IDF and BoW are the most frequent feature extraction used by researchers. This is because both techniques are remarkably simple to understand and implement and offer a lot of flexibility for customization on specific text data.

Table 1: Summary of feature extraction used

References	Term Frequency-Inverse Document Frequency	Bag-of-Words	N-gram	Count Vectorizer	Information Gain
[1]	✓	✓			
[2]	✓	✓	✓		✓
[3]	✓				
[4]		✓			✓
[5]	✓			✓	
[6]	✓		✓		
[7]	✓				
[8]			✓		

2.1.1. ML Classifier

In this article by (Kastrati et al., 2020), four different traditional ML approaches are used to perform sentiment classification about the COVID-19 pandemic. These models are SVM, Decision Tree (DT), RF, and Naive Bayes (NB). (Aljabri et al., 2021) employed six classification algorithms which are Logistic Regression (LR), NB, K-Nearest Neighbor (KNN), Extreme Gradient Boost (XGB), SVM, and RF to perform classification on the collected and pre-processed two separate datasets on remote learning in the Arabic language pertinent to the Saudi Arabian region, each with a big dataset size of over 70,000 tweets and 92,000 tweets.

Following that, (Amin et al., 2021) and (Malla & P.J.A., 2021) used the same three ML approaches as (Kastrati et al., 2021), however, (Malla & P.J.A., 2021) did not use NB, and (Amin et al., 2021) used an additional model which is LR to perform sentiment classification. The suggested machine learning algorithm assist (Amin et al., 2021) in classifying Twitter posts into four groups: confirmed, fatalities, recovered, and suspected.

Furthermore, (Ali, 2021) employed Multinomial Naive Bayes (MNB), LR, KNN, NB, and SVM for classification on two separate Covid-19 datasets about online learning for the experiment. The gathering of data is restricted to "Arabic" tweets only. (Aribowo et al., 2020) also used five distinct types of models which are NB, SVM, DT, RF, and Extra Tree Classifier (ETC). Other than that, five typical ML methods were performed and compared which are KNN, LR, SVM, DT, and RF by authors (Zhang et al., 2020). (Khan et al., 2021) employed RF, Gradient Boosting (GB), ETC, LR, and SVM models whereas in the paper by (Sajib et al., 2019), three different ML algorithms were utilized for classification which is NB, SVM, and LR. (Sajib et al., 2019) gathered 3600 English-language tweets from Twitter using Twitter API and labelled the dataset as negative or positive to analyze the sentiment of Pathao users' tweets.

(Binsar & Mauritsius, 2020) employed three models in this paper which are RF, SVM, and NB on the retrieved 31,003 tweets in Indonesian by scraping posts on Twitter using the selected keywords. Aside from that, numerous different ML

algorithms were used in this study by authors (Jagdale & Deshmukh, 2020), including LR, DT, KNN, RF, SVM, GB, and NB. Lastly, the proposed work by (Karthika et al., 2019) is using the RF and SVM techniques on the data set from the Kaggle website.

From Table 2, we can conclude that RF and SVM are the first most frequent ML used by researchers. The second most frequent are NB and DT. The difference between the first and second most frequent models is RF and SVM provide high accuracy and performance whereas NB and DT have the weakest performance and low accuracy.

Table 2: Summary of machine learning classifier used

Machine Learning Classifier	References											
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
Random Forest	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Support Vector Machine	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Extreme Gradient Boost		✓										
Logistic Regression		✓	✓	✓		✓	✓	✓			✓	
Multinomial Naive Bayes				✓								
Extra Tree Classifier					✓		✓					
Decision Tree	✓		✓		✓	✓			✓		✓	
Naive Bayes	✓	✓	✓	✓	✓			✓		✓	✓	
K-Nearest Neighbor		✓		✓		✓					✓	
Gradient Boosting							✓					✓

2.1.2. Performance Metrics

According to the article by (Kastrati et al., 2021), they examined the data using weighted precision, recall, and F1-Score. According to the data, RF outscored all CML classifier methods with an F1 score of 70.49% and 71.44%, respectively, utilizing TF and TF-IDF. Following that, all the model performances by the authors of the papers (Aljabri et al., 2021), (Zhang et al., 2020), (Khan et al., 2021), and (Malla & P.J.A., 2021), were evaluated using the four standard performance evaluation metrics. According to the findings by (Aljabri et al., 2021), the best

accuracy with a value of 0.899 was obtained with a model that used a 1-gram and TF-IDF as feature extraction approaches, and LR as the model. (Zhang et al., 2020) testing results show that the RF model applying the 1-gram feature extraction method outperformed the other models. According to the study by (Khan et al., 2021), TF-IDF features can improve the performance of supervised ML models, and in this work, the GB surpasses the others and achieves a high accuracy of 96% when paired with TF-IDF features. Other than that, according to the data by (Malla & P.J.A., 2021), the RF model had the best accuracy of 82.06% and precision of 81.92%, but the SVM model outperformed in F1-Score with 82.71% and recall with 84.19%.

Also, the paper by (Amin et al., 2021), employed the four standard evaluation measures in addition to the Confusion Matrix. The classifiers with the highest accuracy, according to the findings, were NB and SVM. (Ali, 2021) assessed the models' performance using precision, recall, and F1-Score. The results show that the suggested model performs well in identifying people's perceptions of coronavirus using an SVM classifier, with a maximum accuracy of 89.65%. In the paper by (Aribowo et al., 2020), accuracy and F-measure were used to assess the model's performance.

According to the findings (Aribowo et al., 2020), the best ML methods in the first and second datasets are ETC and RF, respectively, while the weakest methods are NB in the first dataset and DT in the second. Both authors, (Sajib et al., 2019) and Jagdale & Deshmukh, evaluate the results only based on accuracy. The testing results by Sajib et al., demonstrate that the SVM classifier is the most efficient method out of the three, with an accuracy rate of 82.3% greater than the other classifiers whereas the experiments by (Jagdale & Deshmukh, 2020) show that the RF has the maximum accuracy of 94.90%, and NB has the lowest accuracy of 45%.

The performance of each method by the authors (Binsar & Mauritsius, 2020) is proven using the Confusion Matrix, as well as the ROC and AUC curves. According to the results, the RF model has the highest accuracy level of up to 89%, followed by SVM at 87% and NB at 68%. Lastly, the study by (Karthika et al., 2019) utilized the four standard evaluation measures with the addition of Confusion Matrix and ROC. According to the data by (Karthika et al., 2019), RF has the highest accuracy (97%) while SVM has the slightly lowest accuracy (92%).

Table 3 summarizes accuracy, precision, recall, and F1-score are the most frequent metrics used by researchers whereas Confusion Matrix, AUC, and ROC are the least used metric.

Table 3: Summary of performance metric used

Performance Metric	References											
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
Accuracy		✓	✓		✓	✓	✓	✓	✓		✓	✓
Precision	✓	✓	✓	✓		✓	✓		✓			✓
Recall	✓	✓	✓	✓		✓	✓		✓			✓
F1-Score	✓	✓	✓	✓	✓	✓	✓		✓			✓
Confusion Matrix			✓							✓		✓
Area Under the Curve										✓		
Receiver Operating Characteristics										✓		✓

2.1.3. Discussion

Based on the analysis of the literature, several of the researchers did not use any feature extraction or selection techniques for the sentiment analysis model. Implementing feature extraction or feature selection will help the sentiment analysis model achieve high accuracy scores. Thus, three popular feature extraction approaches will be used in this study: TF-IDF, N-gram, and BoW.

Furthermore, it is reasonable to conclude that most prior studies used RF and SVM. It can also be noted that both RF and SVM provide higher accuracy and performance when compared to other models. The four commonly used assessment metrics based on prior studies will be used to evaluate the models. Additional metric evaluations, such as ROC and AUC, will be used, as both will provide an overall picture of the model's adequacy. As a result, for this paper, the sentiment analysis model will be built using the two ML models RF and SVM and compared with LSTM.

3. Research Methodology

This section describes in detail the research methods used for this paper. Fig. 1 depicts the general flow of the experiment in Section 3.1. The experiment's dataset collection and data pre-processing techniques applied for data cleaning are discusses in Section 3.2 and 3.3. The labelling process was detailed in Section 3.4. Section 3.5 discussed the feature extraction strategies employed in this experiment, while Section 3.6 discusses the model-building process. Lastly, Section 3.7 discusses the procedure of performance evaluation metrics.

3.1. Proposed Methodology

Fig. 1 shows the proposed methodology. The workflow begins with the extraction of data from Twitter into the COVID-19 tweets dataset. After cleaning the dataset using several pre-processing methods, the data is annotated using a lexicon-based method with appropriate sentiment labels. The labelled dataset is then subdivided into training and testing sets for ML models to train and test. In this case, the BoW, N-gram, and TF-IDF feature extraction approaches are employed. After each experiment method's model has completed, the four standard performance evaluation metrics, as well as a classification report, will be generated to analyze the model's performance. The AUC curve is then computed, and the ROC curve is plotted.

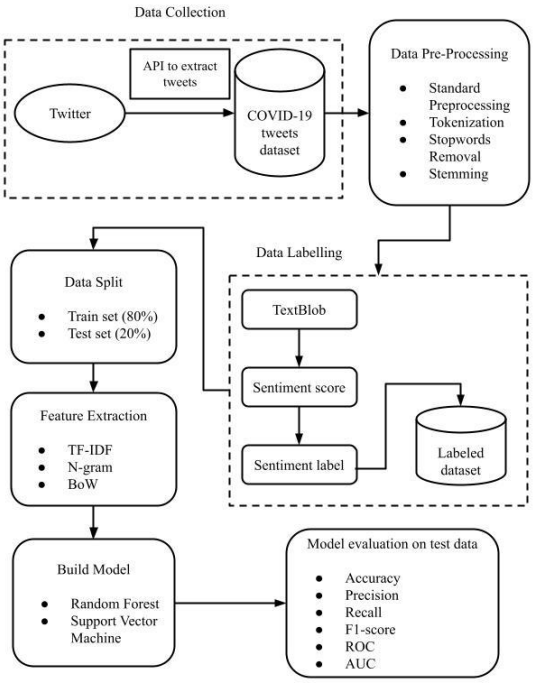

Fig. 1: Proposed methodology

3.2. Data Collection

The API credentials keys received to access the Twitter API are saved in a file and used for authentication after acquiring Twitter authorization to have a developer account. Next, a Standard API search query is created using the selected keywords which are "#covid", "#coronavirus outbreak", "#coronavirus", "#corona", "#corona", "#covid19", "#pandemic", "#social distancing" and "#lockdown". A Tweepy function is used to collect tweets. The filter tweets are based on the above-mentioned search terms, and the language is set to English in this experiment. The extended tweet mode is used to load the entire text of a tweet that would otherwise be truncated. Finally, a query for 1000 tweets is asked because requests for more

than 1000 tweets per minute from Twitter are likely to be rate limited. To acquire over 1,000 tweets for this experiment, the request is executed, and the tweets are stored daily for a total of 8 days, resulting in at least an 8000-tweet dataset. Lastly, the tweet text is extracted and place in a data frame.

3.3. Data Pre-processing

3.3.1. Standard Pre-processing

The data are cleaned by removing mentions, hashtags, hyperlinks, URLs, special characters, numbers, and punctuation marks from the tweets. All the characters are removed from the tweet text because they have no effect, weight, or significance in sentiment analysis. After that, convert all the tweet text to lower case to avoid case sensitive issue. Lastly, any emojis or emoticons are replaced with the text they represent as it is importance in conveying a sentiment.

3.3.2. Tokenization and stop words removal

The stop words in the text are tokenize first before being removed by using the NLTK collection of English stop words. The keywords used in the search query are added to the extended stop words list for this paper. This allows the words in the search query to be isolated from the text data. This is beneficial since it minimizes the complexity of the text data as well as the size of the dataset, which speeds up the training process.

3.3.3. Stemming

The process of simplifying the words from their base form is known as stemming. By lowering the complexity of words, the model can fully understand the meaning of the text. The stemmer in this paper is based on the NLTK library's function.

3.4. Data Labelling

3.4.1. TextBlob

TextBlob is a vocabulary method that can be used for a variety of NLP functions such as sentiment analysis, paraphrase mining, sorting, and so on. The TextBlob sentiment function returns a polarity score ranging from 1 to -1. Tweets with polarity scores less than 0 are considered negative, tweets with polarity scores equal to zero are considered neutral, and tweets with polarity scores greater than zero are considered positive. Subjectivity is expressed by a number between 0 and 1 and indicates whether the statement is considered to be more factual, or opinion based.

For this paper, the TextBlob function is used to get the sentiment subjectivity and polarity. After that, a function to add sentiment label based on its polarity score is created. This paper aims to build a binary classification task rather than a more complex multiclass classification task for the purpose of simplicity. Thus, sentiment

scores equal to zero and sentiment scores greater than zero are combined and labelled as positive. Lastly, map the target variable according to the sentiment label.

A bar chart is plotted to visualize the sentiment distribution in the dataset. As shown in Fig. 2 below, there are more tweets with a positive sentiment than those with a neutral or negative sentiment. There are 1270 positive sentiment, and 696 negative sentiments.

As there is more data with positive sentiment labels than negative sentiment labels, the model will be better at predicting positive sentiments than negative ones. Thus, to balance the dataset, the same number of positive tweets is randomly select and used as negative tweets. This technique is used to balance the dataset so that the prediction model is not biased toward one class over another. This is significant because if the dataset is not balanced, the model will learn to predict classes that are overrepresented in the dataset.

After balancing, the dataset is separated into 80% training set and 20% testing set. This split was chosen because it provides as much data as possible to learn from while still providing a reliable enough amount of data to evaluate the model. The more data there is for the model to learn from, the higher the possibilities of making more accurate predictions on unseen data. If a training set that was too small was employed, the machine would learn the details of the smaller dataset. Thus, an 80/20 training/testing split is a good rule of thumb. The training set will be used for model building, followed by testing set for the test model.

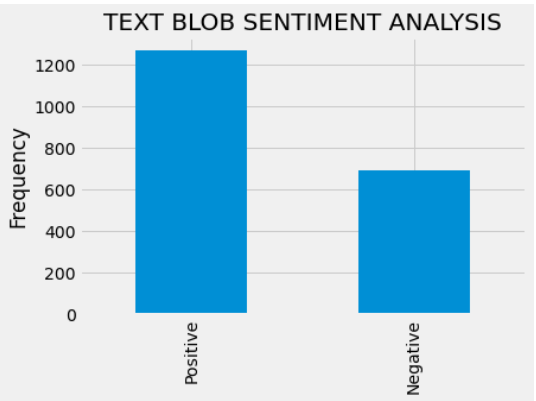

Fig. 2: Sentiment distribution of the tweets

3.5. Feature Extraction

Three different feature extraction methods are implemented for this paper, which are BoW, TF-IDF with 1-gram, and TF-IDF with 2-gram. In this context, the functions from the NLP library are used to construct BoW and TF-IDF features. The following section discusses the feature extraction methods of TF-IDF, N-grams, and BoW.

3.5.1. TF-IDF

TF-IDF is a well-known method for determining the relevance of a word in a document for information retrieval and NLP. The purpose of TF-IDF is to determine the number of times words appear in a large document database. Concisely, TF-IDF is a feature extraction technique that uses weighted features to extract them from textual data. It offers the weight of each phrase in the corpus to help learning models perform better. Smaller TF-IDF values indicate common phrases in the text document, meaning that they are unimportant. Larger TF-IDF values, on the other hand, indicate fewer frequent terms in the text document and are thus significant.

3.5.2. N-gram

The N-gram selection of features and processing approach is widely used in information retrieval and NLP. The N-gram approach is used to save the context of the collected phrases. As mentioned in (Pano & Kashef, 2020) [13], it employs a collection of consecutively organized words based on the value of an N variable. The N-gram is not a textual representation, but it can be used as a feature to describe a text. We employed the 1-gram and 2-gram approaches to representing the context of the Twitter data in this work.

3.5.3. BoW

A technique known as BoW is used to extract features from abbreviated words or information. The BoW is used to count how many times each term occurs, calculate the document's keywords based on the frequency of each word, and produce a frequency histogram from it. Briefly, the BoW is used to increase the lexicon of all unmatched phrases and train models depending on their frequency.

3.6. Classification Model

Two machine learning models, RF and SVM, are implemented in this paper. Section 3.6.1 and Section 3.6.2 discusses the RF and SVM classification models.

3.6.1. RF

RF technique generates many categorizations of DT. Briefly, accordant with (Ankit & Saleena, 2018), RF is a classic ML model based on an ensemble tree since it consists of many DTs that work together as a group. When new data must be categorized, it will traverse the entire forest. Each tree allocates a class to the data, with the greatest occurrence class being picked as the predicted class for the input data.

One of the advantages of RF is that it introduces unpredictability into the model by building many trees and dividing edges using the joint distribution among a random selection of variables chosen at each node. RF uses the Gini index as an input parameter when calculating the defilement of an attribute in terms of classes.

One category (pixel) is randomly selected and stated to correlate to some categories for a given training set x.

3.6.2. SVM

Based on (Ahuja et al., 2019), SVM is an ML technique that is widely used for classification and regression problems. The SVM organizes data into separate groups by locating a state line boundary, also known as a hyperplane, which divides the data set into groups. A certain class is associated with the state line boundary between vectors.

The linear kernel SVM was employed in this study. This technique seeks the best separation function (hyperplane) for separating opinion data into various categories, also known as binary classes in this context. Furthermore, one method for determining the ideal hyperplane is to first determine the outermost data in the two classes that are on the border, and then determine the best hyperplane while taking the outside data into account.

3.6.3. Model Building

Each model is created using a combination of the three different feature extraction methods that were previously implemented. The parameter settings for each model are shown in Table 4 below.

Table 4: Parameter setting for machine learning models

Classifier	Parameters
Random Forest	n_estimators=300, max_depth=300, criterion='entropy', random_state=27
Support Vector Machine	loss='log', l1_ratio=0.15, max_iter=300, n_jobs=4, random_state=101

Four parameters are defined for the RF model. The ‘n estimator = 300’ specifies that RF constructs 300 decision trees participate in the prediction process. The parameter ‘max depth = 300’ limits forest growth to a maximum of 300 levels, significantly decreasing the complexity of the decision tree. The ‘criterion = entropy’ picks the optimal characteristics to partition the data. The ‘random state = 27’ specifies the random seed provided to each tree estimator at each boosting cycle. In this experiment, a value of 27 was utilized because it has been shown to function well for classification models.

Five parameters are defined for the SVM model. The ‘loss = log’ parameter defines the function that will be used to compute the model's loss, which is useful for minimizing loss and improving accuracy. The ‘l1 ratio=0.15’ parameter produces models that are slightly less centralized but produce comparable results. The ‘max_iter=300’ setting limits the maximum number of iterations to 300 predictors and a max depth of 300. The ‘n jobs=4’ defines how many CPUs will be

used for training. The 'random state = 101' specifies that models will be trained four times with 101 different random seeds. The value 101 is utilized because it is a low value that can prevent overfitting.

Following that, the LSTM model is also developed to compare the performance of machine learning models. The model is trained for 6 epochs with batch size=16. The testing dataset is used to assess the model's performance. Fig. 3 show the summary of the model:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 33, 64)            128000

 lstm (LSTM)                 (None, 16)                5184

 dense (Dense)               (None, 1)                 17

=================================================================
Total params: 133,201
Trainable params: 133,201
Non-trainable params: 0
_________________________________________________________________
None
```

Fig. 3: Architecture of the LSTM model

3.7. Performance Evaluation Metrics

After each experiment method's model has been completed, the four standard performance evaluation metrics, as well as a classification report, will be generated to analyse the model's performance. The AUC curve is then computed, and the ROC curve is plotted. The following section discusses evaluation metrics which include four standard evaluation metrics, AUC, and ROC.

3.7.1. Accuracy

As mentioned in this paper (Narasamma & Sreedevi, 2021), accuracy measures the algorithm's efficiency in predicting true values.

3.7.2. Precision

Precision is used to compare the purity of the anticipated True Positive (TP) to the TP of the ground truth.

3.7.3. Recall

Recall assesses the completeness of the predictions by calculating the true positive captured by the model versus the actual true positive.

3.7.4. F1-Score

As state in this paper (Gupta et al., 2021), the F1-score is the proportional mean of precision and recall.

3.7.5. AUC

AUC according to (Satu et al., 2021), is used to investigate ML models by considering the TP and True Negative (TN) rates, which indicate how well positive and negative classes are differentiated.

3.7.6. ROC

The ROC curve is a graph in which the false- positive rate is plotted on the X-axis and the true-positive rate is plotted on the Y-axis.

4. Results and Discussion

This section examines the performance of machine learning models employing the features BoW, 1-gram TF-IDF, and 2-gram TF-IDF. The findings of the model with BoW were shown in Section 4.1. The results of the model using TF-IDF with 1-gram and 2-gram were shown in Section 4.2. The findings of the LSTM model were shown in Section 4.3. Lastly, Section 4.4 discusses the overall findings. In the results table below, classes 0 and 1 represent negative and positive sentiments, respectively.

4.1. Results with BoW

Table 5 shows the BoW results for both models. RF has the highest score in all four measures approximating 98%. Furthermore, SVM achieves slightly lower score in all four measures approximating 96%. Although both models performed well with BoW, RF with BoW performance was superior, more reliable, as seen by all evaluation measures.

Table 5: Model's performance using bag-of-words

Model	Accuracy	Class	Precision	Recall	F1-Score
Random Forest	0.98	0	0.99	0.97	0.98
		1	0.97	0.99	0.98
		Macro Avg	0.98	0.98	0.98
		Weighted Avg	0.98	0.98	0.98
Support Vector Machine	0.96	0	0.95	0.97	0.96
		1	0.97	0.94	0.96
		Macro Avg	0.96	0.96	0.96
		Weighted Avg	0.96	0.96	0.96

4.2. Results with TF-IDF

The features were extracted using varied sizes of N-gram (1-gram and 2-gram) in conjunction with the TF-IDF technique and evaluated using two classification algorithms (RF and SVM). Tables 6 and 7 show the results of research employing

TF-IDF with 1-gram and 2-gram for both classification algorithms in predicting the sentiments of tweets.

RF with 2-gram TF-IDF produced the highest values of 96%, 100%, 100%, and 98%, in all four measures respectively. Furthermore, 1-gram SVM obtains slightly lower results in all four measures with values of 96%, 95%, 95%, and 96%, respectively. Although both models achieved superior performance outcomes with 1-gram and 2-gram samples, RF with 2-gram performance was superior and dependable as seen by all performance indicators.

Table 6: Model's performance using 1-gram term frequency-inverse document frequency

Model	Accuracy	Class	Precision	Recall	F1-Score
Random Forest	0.98	0	0.98	0.97	0.98
		1	0.97	0.98	0.98
		Macro Avg	0.98	0.98	0.98
		Weighted Avg	0.98	0.98	0.98
Support Vector Machine	0.96	0	0.95	0.96	0.96
		1	0.96	0.96	0.96
		Macro Avg	0.96	0.96	0.96
		Weighted Avg	0.96	0.96	0.96

Table 7: Model's performance using 2-gram Term Frequency-Inverse Document Frequency

Model	Accuracy	Class	Precision	Recall	F1-Score
Random Forest	0.98	0	1.00	0.96	0.98
		1	0.96	1.00	0.98
		Macro Avg	0.98	0.98	0.98
		Weighted Avg	0.98	0.98	0.98
Support Vector Machine	0.97	0	0.97	0.96	0.97
		1	0.96	0.97	0.97
		Macro Avg	0.97	0.97	0.97
		Weighted Avg	0.97	0.97	0.97

4.3. Results with LSTM

Table 8 shows the results of the tests that used the LSTM model to predict the sentiments of tweets. The results for LSTM are comparable to those of RF and SVM. The obtained accuracy is comparable to the SVM model. Furthermore, LSTM achieved a similar high performance as both RF and SVM in all four measures.

From this, it is evident that all experiments conducted employing both machine learning models with selected feature extraction are as dependable as LSTM, which is considered among the most preferred models in the field of deep learning.

Table 8: Model's performance using long short-term memory

Model	Accuracy	Class	Precision	Recall	F1-Score
Long Short-Term Memory	0.96	0	0.96	0.97	0.96
		1	0.96	0.95	0.96
		Macro Avg	0.96	0.96	0.96
		Weighted Avg	0.96	0.96	0.96

4.4. Overall Findings

As a result of the overall findings, the RF model outperforms the SVM model for each of the three separate feature extractions. The best model with the best performance is the RF with TF-IDF concatenate with 2-gram model, which consistently achieves the highest score in all four performance evaluation measures with a value greater than 98%.

According to the table, both models with all three different feature extraction achieve high results in all four performance evaluation metrics. This demonstrates that the employed machine learning models have been significantly proven to provide higher performance for sentiment analysis.

Furthermore, the performance of each of these models can be determined with certainty using the ROC curve and AUC value displayed in Fig. 4 and Fig. 5. All models produce a result close to one, indicating a greater TPR, a lower FPR, and a good threshold. The RF model outperforms the SVM model, indicating once again that the RF model is the best performer.

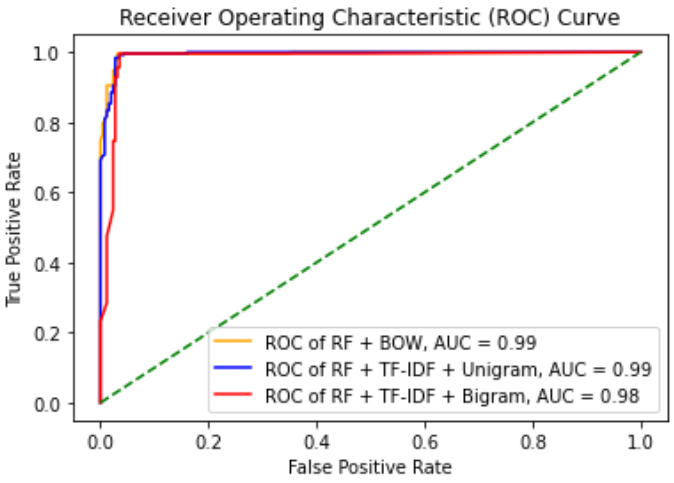

Fig. 4: ROC curves and AUC for RF

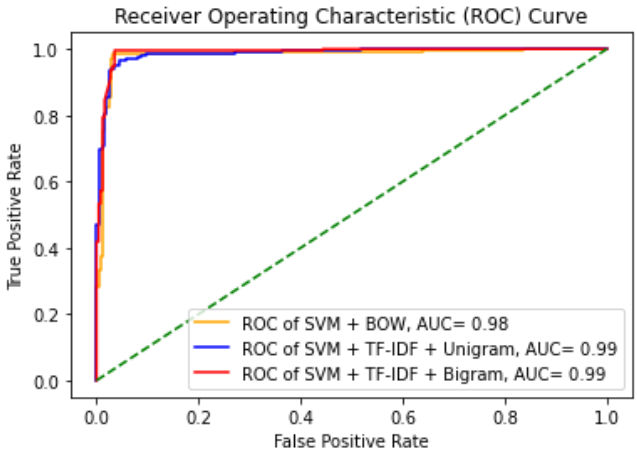

Fig. 5: ROC curves and AUC for SVM

5. Conclusion

One of the drawbacks of this supervised machine learning method is that it can only be utilized for binary classification tasks. A different classifier method and parameters is required to perform multiclass classification for future studies. Furthermore, another limitation for this project is that there were no hyperparameter tuning applied during the building of the classification models. Thus, hyperparameter tuning can be employed in the future studies to find the best parameter that can provide the highest accuracy for each classifier.

Apart from that, this project only presents data from a single train test ratio of 80/20. As a result, multiple sets of train test ratio results, such as 70/30 ratio or 60/40 ratio, can be done for future studies, and the depicted results can be presented for comparison. Moreover, for future work, a larger dataset can be used to further evaluate the model's performance validity as this project only used a small English language tweets dataset. If both the test and training sets were larger, the accuracy may be improved. Aside from that, an additional classifier and feature extraction technique can be utilized to compare the findings with the machine learning algorithms employed in this project along with other tokenization and pre-processing techniques can be opted for better results. Overall, the analysis in this study reveals that the RF with the TF-IDF concatenate with 2-gram features extraction approach is the best performer.

Altogether, this model has the potential to assist the administration and policymakers handle the situation by enabling the development of improved pandemic-fighting measures that take human responses and behavior into consideration. This data was retrieved from positive, neutral, and negative tweets and identified high-frequency information features conveyed and commented on as a response to the pandemic state. To summarize, the pandemic has illustrated the

social and technical constraints of communicating and working in a large-scale, crisis-oriented setting. This compilation of positive, neutral, and negative opinions posted on Twitter throughout the pandemic era can provide a unique perspective on happenings within the Twitter community as well as the issues that COVID-19 poses to society.

References

Alammary, A. S. (2021). Arabic questions classification using modified TF-IDF. IEEE Access,9, 95109–95122. DOI: 10.1109/access.2021.3094115

Aribowo, A. S., Basiron, H., Herman, N. S., & Khomsah, S. (2020). An evaluation of preprocessing steps and tree-based ensemble machine learning for analysing sentiment on Indonesian youtube comments. I*nternational Journal of Advanced Trends in Computer Science and Engineering*, 9(5), 7078–7086. DOI:10.30534/ijatcse/2020/29952020

Aljabri, M., Chrouf, S. M. B., Alzahrani, N. A., Alghamdi, L., Alfehaid, R., Alqarawi, R., Alhuthayfi, J., & Alduhailan, N. (2021). Sentiment analysis of arabic tweets regarding distance learning in Saudi Arabia during the COVID-19 pandemic. *Sensors*, 21(16), 5431. DOI: 10.3390/s21165431

Ali, M. M. (2021). Arabic sentiment analysis about online learning to mitigate covid-19. *Journal of Intelligent Systems*, 30(1), 524–540. DOI:10.1515/jisys-2020-0115

Ankit & Saleena, N. (2018). An ensemble classification system for twitter sentiment analysis. *Procedia Computer Science*, 132, 937– 946. DOI:10.1016/j.procs.2018.05.109

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, 152, 341–348. DOI: 10.1016/j.procs.2019.05.008

Amin, S., Irfan Uddin, M., H. Al-Baity, H., Ali Zeb, M., & Abrar Khan, M. (2021). Machine learning approach for COVID-19 detection on twitter. *Computers, Materials & Continua*, 68(2), 2231– 2247. DOI: 10.32604/cmc.2021.016896

Binsar, F. & Mauritsius, T. (2020). Mining of social media on covid-19 big data infodemic in Indonesia. *Journal of Computer Science*, 16(11), 1598–1609. DOI:10.3844/jcssp.2020.1598.1609

Gupta, P., Kumar, S., Suman, R. R., & Kumar, V. (2021b). Sentiment analysis of lockdown in india during COVID-19: A case study on twitter. *IEEE Transactions on Computational Social Systems*, 8(4), 992–1002. DOI:10.1109/tcss.2020.3042446

Jagdale, R. S. & Deshmukh, S. S. (2020). Sentiment classification on twitter and zomato dataset using supervised learning algorithms. *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*. DOI:10.1109/icsidempc49020.2020.9299 582

Karthika, P., Murugeswari, R., & Manoranjithem, R. (2019). Sentiment analysis of social media network using random forest algorithm. *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. DOI: 10.1109/incos45849.2019.8951367

Khan, R., Rustam, F., Kanwal, K., Mehmood, A., & Choi, G. S. (2021). US based COVID-19 tweets sentiment analysis using TextBlob and supervised machine learning algorithms. *2021 International Conference on Artificial Intelligence (ICAI)*. DOI: 10.1109/icai52203.2021.9445207

Kastrati, Z., Ahmedi, L., Kurti, A., Kadriu, F., Murtezaj, D., & Gashi, F. (2021). A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*, 10(10), 1133. DOI: 10.3390/electronics10101133

Malla, S., & P.J.A., A. (2021). COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets. *Applied Soft Computing*, 107, 107495. DOI: 10.1016/j.asoc.2021.107495

Narasamma, V. L. & Sreedevi, M. (2021). Twitter based data analysis in natural language processing using a novel Catboost recurrent neural framework. *International Journal of Advanced Computer Science and Applications*, 12(5). DOI: 10.14569/ijacsa.2021.0120555

Pano, T. & Kashef, R. (2020). A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. *Big Data and Cognitive Computing*, 4(4), 33. DOI: 10.3390/bdcc4040033

Sajib, M. I., Mahmud Shargo, S., & Hossain, M. A. (2019). Comparison of the efficiency of machine learning algorithms on twitter sentiment analysis of Pathao. *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. DOI: 10.1109/iccit48885.2019.9038208

Satu, M. S., Khan, M. I., Mahmud, M., Uddin, S., Summers, M. A., Quinn, J. M., & Moni, M. A. (2021b). TClustVID: A novel machine learning classification model to

investigate topics and sentiment in COVID-19 tweets. *Knowledge-Based Systems*, 226, 107126. DOI: 10.1016/j.knosys.2021.107126

Zhang, X., Saleh, H., Younis, E. M. G., Sahal, R., & Ali, A. A. (2020). Predicting coronavirus pandemic in real-time using machine learning and big data streaming system. *Complexity*, 1–10. DOI: 10.1155/2020/6688912