Hindawi
Advances in Multimedia
Volume 2023, Article ID 9629700, 12 pages
https://doi.org/10.1155/2023/9629700

*Research Article*

# COVID-19 Infodemic in Malaysia: Conceptualizing Fake News for Detection

**Chee Kuan Lim**,[1] **Zurinahni Zainol**,[1] **Bahiyah Omar**,[2] **and Noor Farizah Ibrahim**[1]

[1]*School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia*
[2]*School of Communication, Universiti Sains Malaysia, Pulau Pinang, Malaysia*

Correspondence should be addressed to Zurinahni Zainol; zuri@usm.my

Received 8 April 2022; Revised 11 March 2023; Accepted 31 March 2023; Published 3 May 2023

Academic Editor: Marco Roccetti



There is an "Infodemic" of COVID-19 in which there are a lot of rumours and information disorders spreading rapidly, the purpose of the study is to build a predictive model for identifying whether the COVID-19 information in the Malay language in Malaysia is real or fake. Under the study of COVID-19 fake news detection, the synthetic minority oversampling technique (SMOTE) is used to generate synthetic instances of real news in the training set after natural language processing (NLP) and before data modelling because the number of fake news is approximately three times greater than that of real news. Logistic regression, Naïve Bayes, decision trees, support vector machines, random forests, and gradient boosting are employed and compared to determine the most suitable predictive model. In short, the gradient-boosting classifier model has the highest value of accuracy and *F*1-score.

## 1. Introduction

This study focuses on detecting COVID-19 fake news in Malaysia. Coronavirus disease (COVID-19) has been announced as a global pandemic by the World Health Organization (WHO) since March 11th, 2020 [1], and there are a lot of discussion and news on social media every day. During the COVID-19 pandemic, people not only need to fight the disease and work on the research to find the cure for COVID-19, but also need to be aware of the growth of the social consequence of the pandemic, which is known as the "Infodemic" of COVID-19. In mid-February 2020, Tedros Adhanom Ghebreyesus, director-general of the WHO, proclaims, "We are not just fighting an epidemic; we are fighting an infodemic" [2]. Based on the definition of the Cambridge Dictionary, infodemic represents a situation that plenty of false information is being spread in a harmful way [3]. The advent of social media has permitted people to share information without restriction. In this digital era, nearly everyone can be exposed to social media, such as Twitter, Facebook, Instagram, and YouTube, and sometimes it is easy to be influenced by fake news without thinking much. In April 2020, the United Nations Communications Response was launched by the United Nations Secretary-General to combat the growth of false information about COVID-19 because disseminating wrong information to the public can not only be harmful to the physical and mental health of the people but also reduce the effectiveness of countries' ability to stop the pandemic [4]. Since, COVID-19 is the first pandemic in history in which technology and social media are employed to keep people informed and connected, it is important to spread the appropriate trust and correct information rather than false and fake information that will jeopardize measures to manage the pandemic [4].

News plays an essential role in our daily lives and society as it is a primary source of information that informs the public's opinion and deliberative processes. Therefore, the public should be receiving real news instead of fake news so that they will not be misled by the information disorder. The council of Europe has introduced a conceptual framework that categorizes information disorder into three different types according to the level of harm associated with it, which

are misinformation, disinformation, and malinformation [5]. Besides these three categories, there are multiple terms defined as categories of fake news in past studies, and the categories of fake news are clickbait, conspiracy theory, fabricated news, hoax, and rumour [6–11].

When COVID-19 has been announced as a global pandemic, many people tend to feel anxious about this pandemic due to uncertainties and the presence of an unknown virus. The situation is getting more serious when people start to create rumours of information disorder and spread rapidly. In Malaysia, the Malaysian Communications and Multimedia Commission (MCMC) encourages the public to visit the official website, which is (https://www.sebenarnya.my) to identify the factuality of the news. Therefore, it is vital to identify the suspicious news related to COVID-19 is real news or fake news, so this study is conducted for COVID-19 fake news detection in Malay language by using machine learning algorithms. Research on COVID-19 misinformation in Malaysian context is very limited and is also one of the motivations to carry out this research study. The research question is as follows: how can COVID-19 fake news in Malay language be detected and predicted by using machine learning algorithms? This led to the objective, which is to build a predictive model for identifying whether the COVID-19 information in Malay language in Malaysia is real or fake.

This study was carried out by collaborating with Malaysia Communications and Multimedia Commission (MCMC), where MCMC provided data collected from social media, such as WhatApp, Facebook, Twitter, and online news platforms. All data are converted to structured data, and conducted predictive data analysis using machine learning algorithms to identify the truth of the news. The outcomes of this study are believed to provide valuable theoretical contributions, and the predictive model of fake news detection is able to play a significant role in reducing systematic cybersecurity risk.

## 2. Related Work

Based on the related work regarding fake news detection, which has been summarized in Table 1, it is clearly shown that the NLP techniques that are commonly used are tokenization, stop word removal, stemming, and lemmatization, whereas the feature extraction that is commonly used is the term frequency-inverse document frequency (TF-IDF). By referring to the machine learning algorithms that have been employed by researchers, there are several algorithms that outperform the other algorithms, such as random forest, linear support vector machines (LinearSVM), decision tree, and logistic regression. Moreover, there are several deep learning algorithms mentioned by the researchers, such as neural networks, Convolutional Neural Network (CNN), Dense Neural Network (DNN), and Passive Aggressive Classifiers that are proven to be better fitted models for fake news detection.

## 3. Materials and Methods

Before further discussion, the summarized methodology of this study is shown in Table 2.

The framework of this study that will be employed is the CrossIndustry Standard Process for Data Mining (CRISP-DM) method, where the steps are shown in Figure 1. First, business understanding is carried out to have a deep understanding of the domain of the study, which is COVID-19 fake news detection. During the business understanding process, the problem statement is identified, and the objective of the study is determined. After that, the data understanding is conducted to identify the characteristics and descriptions of the dataset. A count plot of fake news and real news is visualized to observe whether the data are balanced or not.

For the data preparation process, data preprocessing and feature extraction will be carried out. During data preprocessing, there are a few steps that will be conducted, including tokenization, URL, and punctuation removal, words with the character removal, number removal, stop word removal, stemming, and lemmatization. After data preprocessing, the frequency distributions of each token of a word and word clouds are visualized. Term frequency-inverse document frequency (TF-IDF) is employed in the feature extraction process. The data are split into two sets, which are the train dataset and test dataset with a ratio 7 : 3.

After the data are ready, the machine learning algorithm techniques that will be employed to detect the fake news of COVID-19 are decision tree, logistic regression, random forest, linear support vector machine (SVM), Naïve Bayes, and gradient boosting classifiers. These algorithms will undergo parameter tuning using grid search. After data modelling, the performance of each algorithm is compared using four types of evaluation metrics, such as accuracy, precision, recall, and $F1$-score to identify the most suitable algorithm for COVID-19 fake news detection.

*3.1. Data Understanding.* The dataset is provided by the Malaysia Ministry of Communications and Multimedia Commission (MCMC). The language used in the data is Malay, and the news is distinguished into four categories, including "Makluman" (information), "Palsu" (fake), "Penjelasan" (explanation), and "Waspada" (awareness). The news period is between January 16, 2020, and December 8, 2021.

*3.2. Data Preparation.* For the data preparation process, data preprocessing and feature extraction are carried out. During data preprocessing, a few steps will be conducted by using natural language processing (NLP) as the dataset consists of text data, including tokenization, lowercase transformation, URLs and punctuation removal, words with characters removal, numbers removal, stop words removal, stemming, and lemmatization. After data preprocessing, frequency

Table 1: Summarized information for the related work on fake news detection.

| Journal author | Natural language processing (NLP) | Machine learning algorithms |
|---|---|---|
| Amer and Siddiqui [12] | Tokenization<br>Stop word removal<br>Stemming<br>Term frequency-inverse document frequency (TF-IDF)<br>Count vectorizer | Random forest<br>Decision tree |
| Patwa et al. [13] | Term frequency-inverse document frequency (TF-IDF) | Decision tree<br>Logistic regression<br>Gradient boost<br>Support vector machines (SVM) |
| Madani et al. [14] | Tokenization<br>Stop words removal<br>URLs and punctuation removal<br>Hashtag extraction<br>Words with character removal<br>Stemming<br>TextBlob library–sentiment analysis | Logistic regression<br>Decision tree<br>Naïve Bayes<br>Support vector machines (SVM)<br>Random forest<br>Gradient boosting<br>Multilayer perceptron (MLP) |
| Elhadad et al. [15] | Text parsing<br>Part of speech (POS) tagging<br>Stop words removal<br>Stemming<br>Term frequency-inverse document frequency (TF-IDF)–unigram, bigram, trigram, N-gram (2 : 3)<br>Word embeddings<br>5-fold cross validation | Decision tree<br>kNN<br>Logistic regression<br>Linear support vector Machines<br>Multinomial Naïve Bayes<br>Bernoulli Naïve Bayes<br>Perceptron<br>Neural network<br>Ensemble random Forest<br>Extreme gradient boosting (XGBoost) |
| Felber [16] | Stop word removal<br>Link removal<br>Lemmatization/stemming<br>Reply removal<br>Lowercase transformation<br>XML entity replacement | Linear support vector Machines<br>Logistic regression<br>Multilayer perceptron<br>Naïve Bayes<br>Random forest |
| Pathwar and Gill [17] | Stop words removal<br>Stemming<br>Word embedding<br>Term frequency-inverse document frequency (TF-IDF) | Gradient boost<br>kNN<br>Multinomial Naïve Bayes<br>Random forest<br>Convolutional neural network (CNN)<br>Dense neural network (DNN)<br>Recurrent neural network (RNN)<br>Random multimodel deep learning–CNN, DNN, RNN |

Table 1: Continued.

| Journal author | Natural language processing (NLP) | Machine learning algorithms |
|---|---|---|
| Bondielli and Marcelloni [11] | Sentiment analysis<br>Opinion mining | Support vector machines<br>Decision tree<br>Random forest<br>Logistic regression<br>Recurrent neural networks (RNN)<br>Convolutional neural networks (CNN)<br>Factor analysis of mixed data (FAMD) |
| Ahmad et al. [18] | — | Voting classifier<br>(1) Logistic regression, random forest, kNN<br>(2) Logistic regression, linear support vector machines, classification and regression tree (CART)<br>Machine learning algorithms<br>Logistic regression<br>Support vector machines (SVM)<br>Multilayer perceptron kNN<br>Ensemble learners<br>Random forest<br>Bagging ensemble classifier<br>Boosting ensemble classifier<br>Voting ensemble classifier |
| Joju and Kammath [19] | — | Logistic regression<br>Naïve Bayes<br>Passive aggressive classifiers<br>Random forest<br>Support vector machines |

Table 2: Summarized methodology of the study.

| Objective of study | Machine learning algorithm | Expected outcome |
|---|---|---|
| To identify the COVID-19 information in Malaysia whether it is real or fake | Decision tree<br>Logistic regression<br>Random forest<br>Linear support vector machines<br>Naïve Bayes<br>Gradient boosting classifier | The most suitable algorithm model for COVID-19 fake news detection will be determined by comparing the model evaluation metrics, such as accuracy, precision, recall, and *F*1-score |


Figure 1: The framework of study (CRISP-DM).

distribution of each token of a word and word clouds are visualized. Term frequency-inverse document frequency (TF-IDF) is employed in the feature extraction process. The data are split into two sets, which are the train dataset and the test dataset, with a ratio of 7 : 3. Thus, there are a total of eleven steps to be conducted and the steps are summarized in Table 3.

To investigate the news distribution of the categories, a frequency bar chart is plotted as shown in Figure 2. Based on Figure 2, there are 391 fake news ("Palsu"), 128 explanation news ("Penjelasan"), 37 awareness news ("Waspada"), and 1 information news ("Makluman").

Based on the categories in column "Kategori", a column "Label" is created. Since the news in the category "Waspada" (awareness) is considered as neutral, the rows of data with the category "Waspada" (awareness) are excluded in the data analysis process. For the column "Label," the news in the category "Makluman" (information), and "Penjelasan" (Explanation) column is labelled as real news, whereas the news in the category "Palsu" (fake) are labelled as real news. The frequency bar chart of the label of news is plotted as shown in Figure 3.

The word cloud of fake news is visualized and shown in Figure 4. The higher the frequency of the token of word in the document, the larger the weight of the word among all words in the document. The most frequent token word in the COVID-19 fake news document is "benar" (true), followed by "tular" (contagious), "dakwa" (claim), "hospital" (hospital), and "berita" (news).

Next, when the data preprocessing has been done, data feature extraction has been conducted using the term frequency-inverse document frequency (TF-IDF) with bigram. Term frequency (TF) measures the frequency of a word that appears in a document, and the formula for term frequency is shown [20], whereas inverse document frequency (IDF) measures the significance of the document in the corpus [21], and the formula of inverse document frequency is shown [20]. Panchai [20] indicates that a high value of term frequency and a low value of inverse document frequency will result in a high weight in term frequency-inverse document frequency (TF-IDF).

$$\begin{aligned} \mathrm{TF}(t) &= \frac{\text{Number of Times Term t Appears in a Document}}{\text{Total Number of Terms in the Document}}, \\ \mathrm{IDF}(t) &= \log_e \frac{\text{Total Number of Documents}}{\text{Number of Documents with Term t in it}}. \end{aligned} \tag{1}$$

Based on Figure 3, it is known that data imbalance problem exists, so this problem is required to be resolved before data modelling is carried out. First, the class distribution of the label for train data is identified, and it is found that there are 278 fake news, and 86 real news in the train data, indicating that the class distribution is not uniform among the class labels. Resampling is one of the techniques that can be used to solve the data imbalance problem, and the resampling method includes undersampling and oversampling. Undersampling is employed when there is enough data to be used because this technique will randomly reduce the size of majority class to match the size of minority class, whereas oversampling is employed when there is insufficient of data, so this technique will increase the size of the minority class to match the number of sizes of the majority class [22].

Since the size of the train data for minority classes, which is real news, is small, there will be an insufficient set data, if the undersampling technique is employed. Therefore, this study will employ oversampling to solve the data imbalance issue, where 192 synthetic real news will be randomly generated by SMOTE to balance the class distribution. Badr stated that the Synthetic Minority Oversampling Technique (SMOTE) is the most common technique to be used [23]. For SMOTE, $k$ nearest neighbors (kNN) is used to generate synthetic instances for real news [24]. SMOTE is used because this technique has the advantage of not duplicating the data points that other oversampling techniques usually implement, but SMOTE is creating synthetic data points [27].

### 3.3. Data Modelling

*3.3.1. Decision Tree.* According to Hunt's algorithm, the decision tree is constructed in a top-down recursive divide-and-conquer manner. Sharma and Kumar defined a decision tree is a flowchart-like tree structure that contains internal node, branches, and leaf nodes. An internal node represents a test on an attribute, a branch represents the test's outcome, and a leaf node represents the class label [26]. There are three types of decision tree algorithms, including C4.5, Iterative Dichotomiser 3 (ID3), and Classification and Regression Trees (CART). There are three types of splitting criteria, including information gain (IG), gain ratio, and Gini index, which are used to evaluate the split quality in the decision tree algorithm [27]. Information gain is conceived from information entropy; gain ratio is the advanced version of information gain, whereas Gini index is the criteria that was developed to optimize the splitting of decision tree. The lower the value of entropy, the better the model, as this indicates that the labels of the data are quite uniform. The higher the gain ratio, the greater the quality of the split in the decision tree

Table 3: Summarized procedures for data preparation.

| Index | Procedure |
|---|---|
| 1 | The columns that require for future analysis are extracted |
| 2 | The presence of null or missing values of the extracted data are examined |
| 3 | Columns "Tajuk" and "Keterangan" are combined for future analysis |
| 4 | The news distribution is plotted based on the categories of the news |
| 5 | Column "label" is added, and the news is labelled into "real" and "fake" |
| 6 | The balance of data is checked |
| 7 | The data are preprocessed by natural language processing (NLP) |
| 8 | Frequency distribution of 30 most common token of words and word clouds for COVID-19 related fake news are generated |
| 9 | Feature extraction of the data is generated by term frequency-inverse document frequency (TF-IDF) with bigram |
| 10 | The data are split into 70% train data and 30% test data |
| 11 | SMOTE oversampling technique is employed to solve the data imbalanced problem |

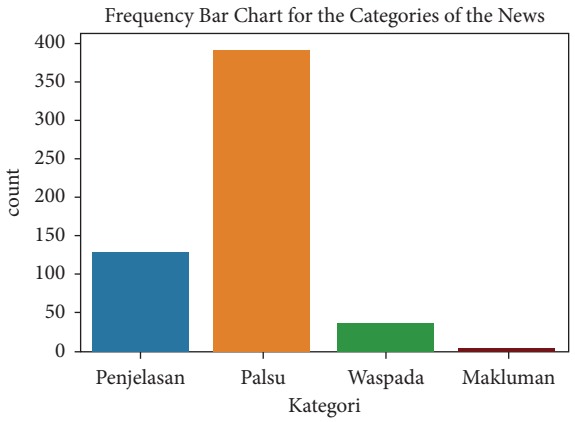


Figure 2: Frequency bar chart of the distribution of news.

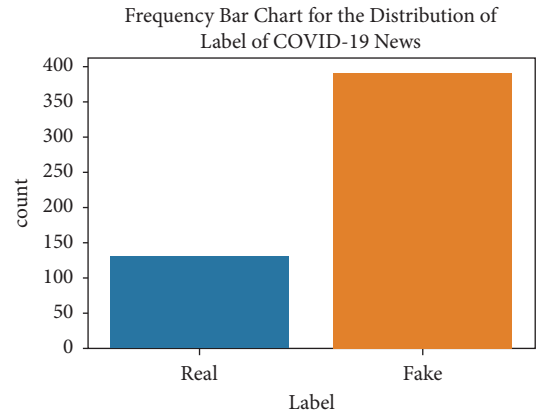


Figure 3: Frequency bar chart of the label of news.

$$\text{Entropy}, H = -\left(\sum p_i \log_2 p_i\right),$$

$$\text{Information Gain}, \text{IG}_{\text{split}} = H - \left(\sum \frac{|D_j|}{|D|} \times H_j\right),$$

$$\text{Intrinsic Information}, \text{II} = -\left(\sum \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|}\right), \quad (2)$$

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Intrinsic Information}},$$

$$\text{Gini}_j = 1 - \left(\sum p_i^2\right),$$

$$\text{Gini}_{\text{split}} = \sum \frac{|D_j|}{|D|} \text{Gini}_j,$$

where $p_i$ is the proportion of class $i$, $D$ is the original dataset, $D_j$ is the $j$th subdataset after being split, $|D|$ and $|D_j|$ are the numbers of samples that belong to the original dataset and the subdataset, respectively, $H_j$ is the entropy of the $j$-th subdataset.


Figure 4: Word cloud of COVID-19 fake news.

*3.3.2. Logistic Regression.* Logistic regression is a regression machine learning algorithm that is employed when the response variable is dichotomous in nature. Logistic regression is a predictive model that is suitable to describe and test hypotheses regarding the relationships between categorical target variable and categorical or continuous independent variables. Due to the binary outcome variable, the shape of the scatterplot is not linear but sigmoidal or S-shaped, which is difficult to describe by linear regression. Therefore, logistic regression is used to solve the nonlinearity problem by applying the logit transformation to the response variable [28].

$$\text{logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x, \tag{3}$$

where $\pi$ = Probability ($Y$ = Outcome of Interest | $X = x$, $a$ specific value of $X$) = $e^{\alpha+\beta x}/1 + e^{\alpha+\beta x}$.

*3.3.3. Random Forest.* Random forest is an ensemble machine learning algorithm that consists of many decision trees. Random forest has the advantage of resolving the overfitting issue caused by decision trees, and the accuracy of random forest will be better than that of decision trees [29]. Moreover, the random forest algorithm can not only handle large datasets with high dimensionality efficiently but also have a lower training time as compared to other algorithms [30]. For a random forest classifier, when the forest has a greater number of trees, it will lead to higher accuracy, and the overfitting problem can be avoided.

*3.3.4. Support Vector Machine (SVM).* Support Vector Machine (SVM), which is a supervised machine learning algorithm, can be distinguished into two types, including linear SVM and nonlinear SVM. A linear SVM classifier is used for binary classes or linearly separable data, whereas a nonlinear data classifier is applicable for nonlinearly separated data [31]. The goal of the SVM algorithm is to identify the best decision boundary, called the optimal hyperplane that can partition the data points in the n-dimensional space [32]. Among the data points, there are some extreme points named support vectors that are used to identify and create the hyperplane. Since the label of the dataset has two classes, the linear SVM algorithm is employed in this study.

*3.3.5. Naïve Bayes.* Naïve Bayes algorithm is a supervised machine learning algorithm based on Bayes theorem that mostly employed in text classification, such as spam filtering and sentiment analysis. Besides that, the Naïve Bayes algorithm is also applicable to high-dimensional training dataset. This algorithm is a probabilistic classifier, which can predict based on the probability of an object [33]. There are three types of the Naïve Bayes algorithms, including Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes. The advantages of Naïve Bayes are easy to build and suitable to be employed in large dataset, so it is outperforming even highly sophisticated classification methods [34]. Naïve Bayes is chosen in this study because this algorithm is suitable to be employed for binary classification problems, and it affords fast, highly scalable model building and scoring [35]. Nelson concluded that multinomial Naïve Bayes outperformed in classification task among the three types of Naïve Bayes algorithms [36].

*3.3.6. Gradient Boosting Classifier.* Gradient boosting classifier is an ensemble machine learning algorithm that combines a group of weak learners, which are the attributes with low accuracy, added together to produce a strong predictive model, which is the model with high accuracy [37]. Gradient boosting is one of the most powerful algorithms, as it can minimize the bias error of the model [38]. There are three elements required for gradient boosting to work, including a loss function, a weak learner, and an additive model. For the gradient boosting framework, the loss function used needs to be optimized as this classifier depends on a loss function. Decision trees are the weak learners that are used in gradient boosting, and the weak learners are added into the additive model over time through a gradient descent procedure to minimize the loss function [39].

*3.4. Model Evaluation.* For supervised machine learning algorithms, model evaluation is carried out to evaluate and compare the models so that the most suitable model for COVID-19 fake news detection is determined. There are four types of evaluation metrics to be used in this study, which are accuracy, precision, recall, and $F1$-score. The calculation of the evaluation metrics is based on the confusion matrix as shown in Table 4 [40].

*3.4.1. Accuracy.* Accuracy represents the ratio of the number of correctly classified data instances to the total number of data instances. The higher the accuracy, the better the fitted model because it indicates that the number of data classified correctly is comparatively higher.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}}. \tag{4}$$

*3.4.2. Precision.* Precision, which is also known as positive predictive value, refers to the ratio of correct positive data instances to the total number of data instances that predicted positive. A good classifier should have a high value of precision, indicating the model can classify the positive data instances correctly

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{5}$$

*3.4.3. Recall.* Recall, which is also known as sensitivity or true positive rate, represents the ratio of correct positive data instances to the total number of positive data instances.

$$\textit{Recall} = \frac{TP}{TP + FN}. \tag{6}$$

*3.4.4. F1-Score.* According to equation (7), it is clearly shown that $F1$ score is highly correlated with precision and recall. Since high precision and high recall are preferred for a good classifier, a model with a high $F1$ score will be chosen to be the more suitable fitted model. Harikrishnan stated that $F1$ score is a better measure compared to accuracy as it is a harmonic mean of precision and recall [40]

Table 4: Confusion matrix.

| | | Predicted label | |
|---|---|---|---|
| | | Negative | Positive |
| True label | Negative | True negative (TN) | False positive (FP) |
| | Positive | False negative (FN) | True positive (TP) |

$$F1\ \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (7)$$

*3.5. Data Science Tool.* For this study, Python is the selected data science tool because the Python Natural Language Toolkit (NLTK) library and the Python scikit learn library can be employed for data exploration, data preparation, and data modelling. Besides that, the language of the dataset is Malay, so Python Malaya documentation is used for data preprocessing. Python is a user-friendly platform that enables the user for text analysis through natural language processing (NLP). Since the NLTK library is designed for text data preprocessing and the scikit learn library consists of different packages, including NumPy, SciPy, Matplotlib, IPython, and Pandas, Python issuitable analytical tools for COVID-19 fake news detection and classification.

## 4. Results and Discussion

*4.1. Models Comparison and Discussion.* Among the six machine learning algorithms, there are two parametric algorithms, two nonparametric algorithms, and two ensemble algorithms. For each algorithm, there are three models carried out to compare the baseline algorithm model, the algorithm model before applying SMOTE, and the algorithm model after applying SMOTE. The baseline algorithm model refers to the model without any parameter tuning, and the parameters are set as default in Python. For the algorithm models before and after SMOTE, parameter tuning has been carried out by using the Python grid search library. Moreover, the training accuracy and testing accuracy of each model are computed to check whether the model is overfitted or not.

The results obtained are recorded in Table 5. Based on Table 5, we found that there are a few models that are overfitted, which achieved 100% accuracy in the testing set. The overfitted models are the logistic regression model and the Naïve Bayes model after applying SMOTE, the decision tree baseline model, the random forest baseline model, and all models in SVM. Among all models, the gradient-boosting classifier performs the best machine learning algorithm because the model is not overfitted and has the highest accuracy and *F*1-score, which are 74.36% and 85%, respectively.

Both parametric machine learning algorithms, which are logistics regression and Naïve Bayes, are overfitted after applying SMOTE. SMOTE increases the real news in the training set from 86 to 278 by generating 192 synthetic real news randomly with the implementation of kNN. Therefore, it is possible that this synthetic real news will be the reason for the model overfitting problem due to multiple samples within the class of real news [41]. To overcome this problem, Kumar suggested using a combination of oversampling and undersampling techniques. Thus, SMOTE and undersampling techniques, such as Tomek or Edited Nearest Neighbour (ENN), can be combined to increase the effectiveness of handling data imbalanced problem in a future study [41].

By comparing the two nonparametric machine learning algorithms used, Table 5 shows that the three models for support vector machine (SVM) are overfitted, whereas for decision tree classifier, the baseline model is overfitted, but the models no longer overfit after hyperparameter tuning by using grid search. Therefore, the model for the decision tree classifier indicates that parameter selection is one of the ways to overcome the overfitting problem, which is also mentioned by IBM Cloud Education [42]. One of the possible reasons that SVM models are overfitted is due to parameter gamma, because Rohilla mentioned that parameter gamma is used to control overfitting in SVM [43]. However, Yidirim mentioned that the overfitting problem will occur when very large gamma values are employed [44], so optimizing gamma values can probably overcome overfitting problem, where the gamma value is suggested to be within the range of 0.0001 to 10.

For ensemble learning algorithms, random forest, and gradient boosting classifiers, there is only one overfitted model, which is the baseline model of random forest classifier that is similar to the decision tree classifier model. Hyperparameter tuning and using ensemble methods have avoided overfitting in random forest and gradient boosting classifier models [42]. Among these two ensemble learning methods, the gradient-boosting classifier model performs better than random forest by comparing the accuracy and *F*1-score because gradient boosting performs better than random forest when the data set is unbalanced [45].

Based on the related work discussed in the related work session, the research studies concluded that random forest, linear SVM, decision tree, and logistic regression outperformed the other machine learning algorithms. However, the best classifier model for COVID-19 fake news detection in this project is a gradient-boosting classifier. The imbalanced size of the project dataset is one of the reasons, because Glen mentioned that a gradient-boosting classifier is a better algorithm model for the imbalanced data [45]. Besides that, gradient boosting classifier is an ensemble model, which is a combination of multiple sets of decision trees that is able to produce better results than a single model [46].

*4.2. Best Model Evaluation.* The best classifier model is the gradient-boosting algorithm model before applying SMOTE. The best parameters for a gradient-boosting classifier are a 0.15 learning rate, a maximum depth of 4, and 20 estimators. Based on Table 5, the training accuracy of the model is 88.19%, which does not deviate too much from the testing accuracy score, suggesting that this model is not overfitted. Based on Figure 5, the accuracy of the classification model is 74.36%, indicating that 116 labels are

Table 5: Summarized results of algorithm models of COVID-19 fake news detection.

| Algorithm | Model | Train score (%) | Test score (%) | Accuracy (%) | Precision (%) | Recall (%) | *F*1-score (%) |
|---|---|---|---|---|---|---|---|
| Logistic regression | Baseline | 76.37 | 72.44 | 72.44 | 72 | 100 | 84 |
| | Before SMOTE | 99.18 | 73.08 | 73.08 | 74 | 96 | 84 |
| | After SMOTE | 100 | 72.44 | 72.44 | 73 | 98 | 84 |
| Naïve Bayes | Baseline | 83.24 | 72.44 | 72.44 | 72 | 100 | 84 |
| | Before SMOTE | 83.24 | 72.44 | 72.44 | 72 | 100 | 84 |
| | After SMOTE | 100 | 73.08 | 73.08 | 78 | 88 | 82 |
| Decision tree | Baseline | 100 | 71.79 | 71.79 | 76 | 88 | 82 |
| | Before SMOTE | 77.75 | 72.44 | 72.44 | 72 | 100 | 84 |
| | After SMOTE | 94.60 | 73.72 | 73.72 | 79 | 88 | 83 |
| Support vector machine | Baseline | 100 | 72.44 | 72.44 | 72 | 100 | 84 |
| | Before SMOTE | 100 | 72.44 | 72.44 | 73 | 98 | 84 |
| | After SMOTE | 100 | 72.44 | 72.44 | 73 | 98 | 84 |
| Random forest classifier | Baseline | 100 | 73.08 | 73.08 | 73 | 100 | 84 |
| | Before SMOTE | 84.89 | 72.44 | 72.44 | 74 | 96 | 84 |
| | After SMOTE | 89.56 | 73.08 | 73.08 | 74 | 98 | 84 |
| Gradient boosting classifier | Baseline | 93.13 | 72.44 | 72.44 | 74 | 96 | 83 |
| | **Before SMOTE** | **88.19** | **74.36** | **74.36** | **75** | **97** | **85** |
| | After SMOTE | 98.38 | 75.64 | 75.64 | 80 | 89 | 84 |

*The most suitable machine learning algorithm model is bold.

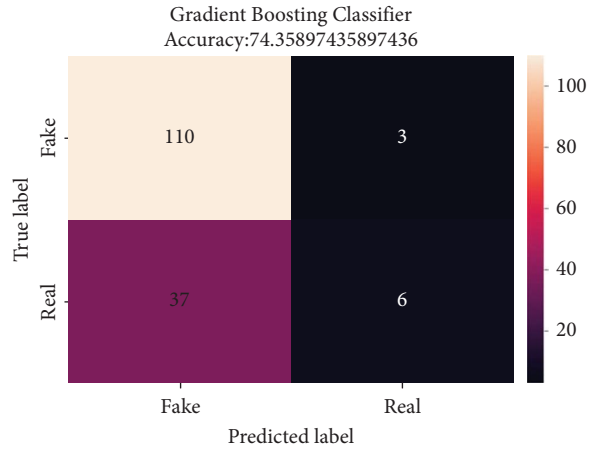


Figure 5: Confusion matrix of a gradient boosting classifier model.

Table 6: Evaluation metrics of the COVID-19 fake news classification.

| Metric | Percentage (%) |
|---|---|
| Accuracy | 74.36 |
| Precision | 74.82 |
| Sensitivity/recall | 97.35 |
| *F*1-score | 84.61 |

| Classification | Report for Gradient Boosting Classifier | | | |
|---|---|---|---|---|
| | precision | recall | f1-score | support |
| Fake | 0.75 | 0.97 | 0.85 | 113 |
| Real | 0.67 | 0.14 | 0.23 | 43 |
| accuracy | | | 0.74 | 156 |
| macro avg | 0.71 | 0.56 | 0.54 | 156 |
| weighted avg | 0.73 | 0.74 | 0.68 | 156 |
| Number of mislabeled points out of a total 156 points : 40, performance 74.36% | | | | |

Figure 6: Classification report for gradient boosting classifier model.

correctly classified out of 156 labels. Since 110 fake news stories are correctly categorized as predicted fake news, the precision of the fake news classification is 74.82%. According to Table 6 and Figure 6, the recall of the fake news classification is higher than the precision, which is 97.35%. The value of the *F*1-score of the gradient-boosting classifier is 84.61%, which is comparatively high.

## 5. Conclusion

This study is regarding COVID-19 fake news detection. Due to the data imbalance problem, SMOTE is used to overcome this problem by generating synthetic data in the training set. However, logistic regression, Naïve Bayes, and SVM are overfitted after applying SMOTE, and it is possible that this synthetic real news will be the reason for the model overfitting problem due to the multiple samples within the class of real news [41]. Therefore, due to overfitting problem and hyperparameter tuning, the gradient boosting classifier model before SMOTE is concluded to be the best classifier model for COVID-19 fake news detection, with the highest accuracy of 74.36% and an *F*1-score of 85%. This result can be supported by the articles written by Glen and Demir (n.d.), as they mentioned that gradient boosting classifier can not only deal with imbalanced data [45], but also produce an improved and better result than a single machine learning model [46]. In future studies, a combination of SMOTE and undersampling techniques, such as Tomek or Edited Nearest Neighbour (ENN), is suggested to be implemented to increase the effectiveness of handling data imbalance problems and as an alternative to solve the model overfitting problem in this study.

## Data Availability

The data used in this study are collected and provided by the Malaysia Ministry of Communications and Multimedia Commission (MCMC). The data can be obtained from the website (https://www.sebenarnya.com).

## Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this study.

## Acknowledgments

The authors are wishing to thank the Malaysia Ministry of Communications and Multimedia Commission (MCMC) for supporting this work by providing data. This research was supported by the School of Computer Sciences and School of Communication, Universiti Sains Malaysia and Malaysia Communications and Multimedia Commission (MCMC).

## References

[1] World Health Organization, "WHO director-general's opening remarks at the media briefing on COVID-19," 11 March 2020, https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19--11-march-2020.

[2] Department of Global Communications, "UN Takles "Infodemic" of Misinformation and Cybercrime in COVID-19 Crisis," 31 March 2020, https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%98infodemic%E2%80%99-misinformation-and-cybercrime-covid-19.

[3] Cambridge University Press, "Cambridge dictionary," 2021, https://dictionary.cambridge.org/dictionary/english/infodemic.

[4] World Health Organization, "Managing the COVID-19 Infordemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation," 23 September 2020, https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation.

[5] Council of Europe portal, "Information Disorder," 2021, https://www.coe.int/en/web/freedom-expression/information-disorder.

[6] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications*, vol. 153, Article ID 112986, 2020.

[7] E. C. Tandoc, J. Jenkins, and S. Craft, "Fake news as a critical incident in journalism," *Journalism Practice*, vol. 13, pp. 673–689, 2019.

[8] J. Shin, L. Jian, K. Driscoll, and F. Bar, "The diffusion of misinformation on social media: temporal pattern, message, and source," *Computers in Human Behavior*, vol. 83, pp. 278–287, 2018.

[9] A. P. Weiss, A. Alwan, E. P. Garcia, and J. Garcia, "Surveying fake news: assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking," *International Journal for Educational Integrity*, vol. 16, pp. 1–30, 2020.

[10] Z. F. Chen and Y. Cheng, "Customer response to fake news about brands on social media: the effects of self efficacy, media trust, and persuasion knowledge on brand trust," *Journal Of Product & Brand Management*, vol. 29, 2019.

[11] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, pp. 38–55, 2019.

[12] A. Y. A. Amer and T. Siddiqui, "Detection of COVID-19 fake news text data using random forest and decision tree classifiers," *International Journal Of Computer Science And Information Security (Ijcsis)*, vol. 18, no. 12, pp. 88–100, 2020.

[13] P. Patwa, S. Sharma, S. Pykl et al., "Fighting an infodemic: COVID-19 fake news dataset," *Communications in Computer and Information Science*, vol. 1402, 2021.

[14] Y. Madani, M. Erritali, and B. Bouikhalene, "Using artificial intelligence techniques for detecting Covid-19 epidemic fake

news in Moroccan tweets," *Results in Physics*, vol. 25, Article ID 104266, 2021.

[15] M. K. Elhadad, K. F. Li, and F. Gebali, "Detecting misleading information on COVID-19," *IEEE Access*, vol. 8, pp. 165201–165215, 2020.

[16] T. Felber, "Constraint 2021: machine learning models for COVID-19 fake news detection shared task," 2021, https://arxiv.org/abs/2101.03717.

[17] P. Pathwar and S. Gill, "Tackling COVID-19 infodemic using deep learning," 2021, https://arxiv.org/abs/2107.02012.

[18] I. Ahmad, M. Yousaf, S. Yousaf, and M. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, Article ID 8885861, 11 pages, 2020.

[19] S. Joju and P. S. Kammath, "Analysis on fake news detection using machine learning," *International Research Journal Of Engineering And Technology (Irjet)*, vol. 8, no. 7, pp. 181–187, 2021.

[20] A. Panchai, "NLP - Text Summarization using NLTK: TF-IDF Algorithm," 10 Jun 2019, https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3.

[21] W. Scott, "TF-IDF from scratch in Python on a real-world dataset, Data Science," 15 February 2019, https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089.

[22] Y. Wu and R. Radewagen, "7 techniques to handle imbalanced data," Jun 2017, https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html.

[23] W. Badr, "Having an imbalanced dataset? Here is How You can Fix It," 22 February 2019, https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb.

[24] C. Paduraria and M. E. Breaban, "Dealing with data imbalance in tect classification," in *Proceedings of the in 23rd international conference on knowledge-based and intelligent information & engineering systems*, January 2019.

[25] J. Korstanje, "SMOTE," *Towards Data Science*, 2021, https://towardsdatascience.com/smote-fdce2f605729.

[26] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," *International Journal of Science and Research (IJSR)*, vol. 5, no. 5, pp. 2094–2097, 2016.

[27] M. P. Tung, "Information gain, Gain Ratio and Gini Index," 4 January 2020, https://tungmphung.com/information-gain-gain-ratio-and-gini-index/.

[28] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The Journal of Educational Research*, vol. 96, pp. 3–14, 2002.

[29] O. Mbaabu, "Introduction to Random Forest in Machine Learning," 11 December 2020, https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

[30] Javatpoint, "Random Forest Algorithm," 2022, https://www.javatpoint.com/machine-learning-random-forest-algorithm.

[31] Javatpoint, "Support vector machine algorithm," 2022, https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[32] R. Gandhi, "Support Vector Machine - Introduction to Machine Learning Algorithms," 8 June 2018, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[33] Javatpoint, "Naive Bayes classifier algorithm," 2022, https://www.javatpoint.com/machine-learning-naive-bayes-classifier.

[34] S. Ray, "6 Easy Steps to Learn Naive Bayes Algorithm with Codes in Python and R," 11 September 2017, https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[35] Oracle Help Center, "Data mining concepts - naive Bayes," 2022, https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_nb.htm?fbclid=IwAR235O4liRqX1epj9aNV8jOjB9jC-C20mjhgoPGoT0uXpLSvq67I3f39ngc#DMCON018.

[36] M. Ismail, N. Hassan, and S. S. Bafjaish, "Comparative analysis of naive bayesian techniques in health-related for classification task," *Journal Of Soft Computing And Data Mining*, vol. 1, no. 2, pp. 1–10, 2020.

[37] D. Nelson, "Gradient Boosting Classifiers in Python with Scikit-Learn," 9 August 2019, https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/.

[38] N. Tarbani, "How the Gradient Boosting Algorithm Works?," 19 April 2021, https://www.analyticsvidhya.com/blog/2021/04/how-the-gradient-boosting-algorithm-works/.

[39] J. Brownlee, "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning," 9 September 2016, https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.

[40] N. B. Harikrishnan, "Confusion matrix, accuracy, precision, recall, *F*1-score," 2019, https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd.

[41] S. Kumar, "Stop using SMOTE to handle all your Imbalanced Data," 3 May 2021, https://towardsdatascience.com/stop-using-smote-to-handle-all-your-imbalanced-data-34403399d3be.

[42] Ibm Cloud Education, "Overfitting," 3 March 2021, https://www.ibm.com/cloud/learn/overfitting#:%7E:text=Overfitting%20is%20a%20concept%20in,unseen%20data%2C%20defeating%20its%20purpose.

[43] A. Rohilla, "A Brief Introduction to Support Vector Machine," 2 November 2018, https://adityarohilla.com/2018/11/02/a-brief-introduction-to-support-vector-machine/.

[44] S. Yildirim, "Hyperparameter Tuning for Support Vector Machines - C and Gamma Parameters," 1 June 2020, https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167.

[45] S. Glen, "Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply," 28 July 2019, https://www.datasciencecentral.com/decision-tree-vs-random-forest-vs-boosted-trees-explained/#:%7E:text=If%20you%20carefully%20tune%20parameters,to%20tune%20than%20random%20forests.

[46] N. Demir, "Ensemble methods: elegant techniques to produce improved machine learning results," 2022, https://www.toptal.com/machine-learning/ensemble-methods-machine-learning#:%7E:text=Ensemble%20methods%20are%20techniques%20that,winning%20solutions%20used%20ensemble%20methods.