Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Sentiment Analysis on Acceptance of COVID-19 Vaccine for Children based on Support Vector Machine

Muhammad Adam Sani Mohd Sofian[1] , Norlina Mohd Sabri [1,*], Ummu Fatihah Mohd Bahrin[1], N.Hrishvanthika[2], Norulhidayah Isa[1]

[1] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Terengganu, Kampus Kuala Terengganu, 21080 Kuala Terengganu, Malaysia
[2] Department of Information Technology, Bharathiar University, Coimbatore 641046 , Tamil Nadu, India

ABSTRACT

Sentiment Analysis is a Natural Language Processing (NLP) branch that focuses on the analysis of the public opinions based on specific topics. The studies on sentiment analysis have been increasing since the COVID-19 pandemic in 2020. After the herd immunity has been reached around the world, the attention has shifted toward the children's COVID-19 vaccination program. It is beneficial to mine people's opinions regarding this issue since kids are often perceived as vulnerable and need the parents' consent. The machine learning based sentiment analysis has proven to be more efficient in the sentiment classification. This study aims to explore the capability of the Support Vector Machine (SVM) algorithm in the sentiment classification of the COVID-19 vaccination for children based on Twitter data. SVM has been one of the powerful algorithms, but never tested in this classification problem. The dataset for this project was scraped using Twitter API Tweepy based on the keywords such as "COVID vaccine children" and "5 to 11 vaccine". The SVM model is based on the Linear Kernel and has been tested with the hold out method. The model has performed better in the balanced dataset with the implementation of oversampling by the SMOTE technique. A GUI prototype has also been developed using TKinter for the SVM classifier. The results have been divided into the data exploratory and the algorithm's performance analyses. In this study, it is found that there are actually more people are supporting the COVID-19 vaccination for children. Meanwhile, based on the performance analysis, SVM has been able to classify the positive and negative tweets with an acceptable accuracy of 82%. The future work includes the scrapping of data from other social media platforms for larger demographics and also to compare SVM performance with other algorithms.

Keywords:
COVID-19 children vaccine; acceptance; sentiment analysis; support vector machine; Twitter data

1. Introduction

Sentiment Analysis is a statistical research method under the "Natural Language Processing" technique that focuses on people's thoughts, opinions and perceptions in which text mining would be utilized [1]. It is also known as opinion mining, which is the process of understanding, extracting

[*] *Corresponding author.*
E-mail address: norli097l@uitm.edu.my

https://doi.org/10.37934/araset.58.2.252270

and processing textual data. It would classify whether the feedback is positive, neutral or negative [2]. The purpose of sentiment analysis is to analyze the opinions obtained from the social media, especially Twitter [3]. Twitter is the top-rated and dynamic platform in data acquisition with a total of around 166 million users daily, consisting of users with a variety of age groups [4]. Twitter can be utilised as a data bank to garner the sentiment to obtain information and identify the actual public emotion [5]. The tweets data can also show the attitudes and reflections of different opinions in different locations [6]. These public opinions usually become a consideration in decision-making in the govenrment and also private organizations. The examples of sentiment analysis utilizations are in news articles, blogs, stock market, political debates, and movie reviews [2].

COVID-19 has been infected millions of people worldwide since 2020 and many measures have been taken ever since by governments [7]. The disease has challenged the global health system, changed the way of life and has claimed many lives [8]. Instead of movement control order and new normal rules, vaccines have also been compulsory for all of the people in the world to take. The adults have been the first to be vaccinated before it was opted to the children. In October 2021, the American Academy of Pediatrics revealed that almost 6.3 million children had tested positive for COVID-19 since the onset of the pandemic. COVID-19 was the eighth-highest killer of kids ages 5 to 11 over the past year [9]. This concerning matter has been a call for action for authorities such as the Food and Drug Administration (FDA) of the United States to allow COVID-19 vaccination shots for children of 5 to 11 years old [10]. The Centers for Disease Control and Prevention (CDC) of the United States had then suggested that every child of age 5 to 11 get their Pfizer shot [11]. As a result, this announcement had received plenty of critical responses from Twitter users, such as parents and guardians. In addition, there were news surrounding the internet and newspapers that Pfizer BioNTech vaccines could caused myocarditis, which is a chest pain burning problem [12]. Furthermore, medical doctors have confirmed that children would likely have the same effects as adults. These headlines may leave the public, especially the parents being hesitant to allow their children to get their COVID-19 shots. This argument was due to the health risk of children getting side effects from the vaccine [13]. This news published by the organization may affect public views regarding this matter. If the reactions are mostly negative, it may become a problem for the governments to enforce the COVID-19 childhood vaccination program not only in the United States but all over the world.

However, since children are vulnerable and there are also benefits from the COVID-19 vaccination, the public should be educated on the importance of the disease immunisation. A study released later revealed that the Pfizer vaccine is 91% effective at preventing symptomatic COVID-19 in children [14]. The United Kingdom Joint Committee on Vaccination and Immunization (JCVI) also agreed that the side-effects from getting vaccination shots are extremely rare and could be resolved fast [13]. Carrying out the sentiment analysis may help governments to decide whether extra steps are needed to convince the public that COVID-19 vaccines are safe for children. The public view is important as the children need adults' permissions to get their vaccines. Therefore, whether or not they get vaccinated is fully depending on their parents and guardians. Acquiring data from Twitter might be more representative of actual opinions than survey data and provide opportunities for real-time analyses of public sentiments [15].

Based on the issue, the sentiment analysis on the acceptance of COVID-19 vaccine for children based on the machine learning technique is proposed in order to gain the public views on the matter. The objective of the study is to explore the capability of the Support Vector Machine (SVM) in the sentiment analysis on the acceptance of the children vaccination based on Twitter data. In public health research, Twitter has become a significant platform for gathering public opinion [6]. The machine learning technique has been chosen as it could outperform the Lexicon-based method [16].

Based on the previous studies, SVM has outperformed other algorithms such as Naïve Bayes, Linear Regression, Decision Tree, Random Forest, and XG Boost [5][17]. It is expected that SVM could also generate good performance in this sentiment analysis study. The results of the study could alert the authorities and also help them to make important decisions related to children vaccination.

The paper is structured as follows: Section I presents the Introduction, while section II discusses the Literature Review, which contains the brief explanations on the Similar Work and Support Vector Machine. Section III discusses about Methodology and Section IV presents the Result and Discussion. Finally, Section V concludes the study.

2. Literature Review

2.1 Similar Works

Most of the sentiment analysis studies on COVID-19 vaccination are not specific with children. There are several works that have been summarized which are related to machine learning based sentiment analysis on COVID-19. Table 1 shows the summary of the similar works in this study.

A study in Indonesia has implemented SVM in the sentiment analysis of the public views on Sinovac and Pfizer vaccines based on Twitter data. In the study, SVM has outperformed Naïve Bayes for the classification problem [5]. Another study has also analyzed the COVID-19 vaccine-related tweets and the reported result has shown that the VADER tool could picked up more negative sentiments than TextBlob [4].

The study conducted by Ezhilan *et al.*, [18] has implemented several machine learning based approach which are Naïve Bayes, Support Vector Machine, Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) for the sentiment analysis. The objective of the study is to analyze the sentiments of users across various countries during the pandemic based on tweet data. In the study, the highest accuracy was obtained by CNN. LSTM and Naïve Bayes have also been applied in the classification problem by Alharbi and Alkhateeb [1]. The study aims to help governments and private organization to understand the public sentiments towards the pandemic. This study yielded that LSTM has outperformed Naïve Bayes.

Several machine learning algorithms have been applied by Banik *et al.*, [17], which are Linear Regression, Multinomial Naïve Bayes, Decision Tree, Random Forest, and SVM. The objective of the stidy is to analyze the emotional status of an individual during a pandemic in order to maintain good mental health. The result of the study has shown that SVM has generated the highest accuracy of 91% and 93.03% both in multiclass and binary classes, respectively [17].

In the study by Jayasurya *et al.*, [19], there were 14 machine learning algorithms that have been applied. The objective is to use Natural Language Processing (NLP) and machine learning algorithms to analyze the sentiment of tweets on the vaccination drive. The result is that the SVM, Random Forest and Logistic Regression have generated good performance among of the 14 algorithms.

The study by Nezhad and Deihimi [20] has applied the deep learning model, which is the CNN-LSTM. The objective of the study is to analyze Iranian's perceptions on the COVID-19 vaccination by comparing the local and international vaccines. The result shows that CNN-LSTM produced good results in the sentiment classification [20].

Based on these similar works, SVM has been chosen as the algorithm has generated good performance in the sentiment analysis classification problems. The advantage of SVM is that it could produce high accuracy classification compared to other algorithms in sentiment analysis [19]. Therefore, the capability of SVM in solving another classification problem could further be explored in this study.

Table 1
Similar works related to the project

No.	Algorithm	Title	Objective	Result	Ref
1.	Naïve Bayes, SVM, Random Forest	Sentiment Analysis on COVID19 Vaccines in Indonesia: From the Perspective of Sinovac and Pfizer	To do a sentiment analysis of the two types of vaccines (“Sinovac”, “Pfizer) on the Twitter platform.	SVM contributes the best precision, recall, and F1 score values (85% accuracy).	[5]
2.	VADER and TEXT-BLOB	Analysing Public Sentiments Regarding COVID-19 Vaccine on Twitter	To analyze COVID-19 Vaccine related tweets.	VADER picked up the negative sentiment more compared to TextBlob.	[4]
3.	NB, SVM, CNN and RNN (LSTM).	Sentiment Analysis and Classification of COVID-19 Tweets	To analyze the sentiments of users across various countries based on tweets posted during the pandemic.	The highest accuracy is achieved by CNN (94.4%) Naïve Bayes and Support Vector Machine produced better accuracy when using unigram.	[18]
4.	LSTM and Naïve Bayes	Sentiment Analysis of Arabic Tweets Related to COVID-19 Using Deep Neural Network	To help different Govern-ment and private organiza-tions to understand public sentiments towards this pandemic.	LSTM model performs better with an accuracy of 99%	[1]
5.	Linear Regression (LR), Naïve Bayes, Decision Tree (DT), Random Forest (RF), XGBoost (XGB), and SVM	Classification of COVID19 Tweets based on Sentimental Analysis	To analyze the emotional status of an individual during pandemic to maintain a good mental health.	SVM outperformed other algorithms with accuracy 91% in multiclass dataset and 93.03% in binary class dataset.	[17]

6.	Decision Tree, SVM, K-Nearest Neighbor, Random Forest, LR, XGBoost and Naïve Bayes.	Analysis of Public Sentiment on COVID-19 Vaccination Using Twitter	To use NLP and machine learning algorithms to analyze the sentiment of tweets on the vaccination drive.	SVM, Random Forest and Logistic Regression generated good accuracy out of 14 algorithms.	[19]
7.	CNN and LSTM.	Twitter Sentiment Analysis from Iran About COVID 19 Vaccine	To analyse Iranian's perceptions on the COVID-19 vaccination and to compare their opinions on the local vs. imported COVID-19-vaccines.	CNN-LSTM produced good results in the sentiment classification.	[20]

2.2 Support Vector Machine with Linear Kernel

The Support Vector Machine (SVM) performs classification by identifying the optimal separating hyperplane, which in 2D can be drawn as a straight line. The straight line is to be in the middle of two sets of the training samples in the feature space, which leads to maximal generalization [21]. The purpose of SVM is to design a decision surface, which separates the margin between the class levels. The algorithm identifies this hyperplane using support vectors and margins. It then divides the space into two half parts, where those data values of different levels are divided [21]. For a better understanding of the algorithm, Figure 1 illustrates the concept of SVM. The solid line is representing the hyperplane that separates between two classes. There are also three points scattered, which can be observed on the dotted lines, which are 2 reds and 1 blue. These points are called Support Vectors, which has the minimum distance to the hyperplane, compared to other data points. The goal of SVM is to maximize the margin.

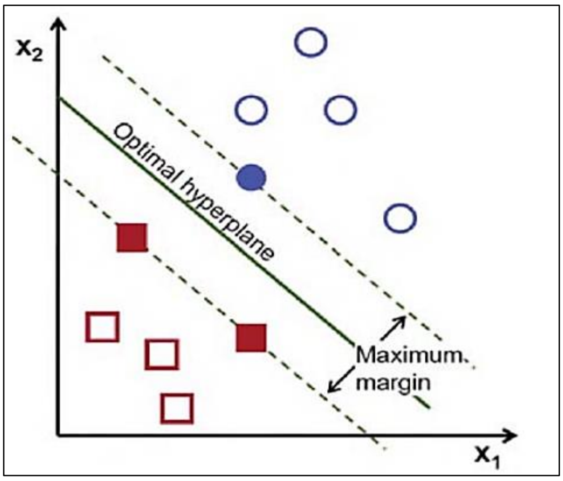


Fig. 1. Data classifications of two classes using hyperplane (Source: [22])

The computations of data points separation leaned on a kernel function. Choosing the suitable kernel function is vital for every application of SVM. There are two types of kernels, which are linear and non-linear kernels. Kernel tricks are utilized to transform the non-linear space into linear space [21]. The linear kernel is the simplest kernel and uses uncomplicated functions. For the linear kernel, it is used when the data is separable. Besides, it is mostly utilized when there are many features in particular data, such as text classification fields [23]. Linear kernel function possesses separable linear, small dataset sample and relatively shortest training duration [21]. In this study, the linear kernel has been chosen for its suitablity with this classification problem.

The non-linear kernels can be categorized into different types, which are the polynomial kernel, Gaussian Radial Basis Function (RBF) and Sigmoid kernel. The Polynomial kernel is suitable to be applied in a smaller dataset [21]. The RBF kernel could normally be used for all types of data and could produce great precision [23]. The Sigmoid kernel, which is derived from the Neural Network field, is often utilized to activate artificial neurons.

In its simplest type, SVM only performs binary classification. For multiclass classification, similar concept is applied after breaking down the multi classification problem into a few binary classification problems [24]. The issue for SVM is that sometimes, the data are not linearly separable, making it difficult to set the hyperplane. Nevertheless, there are have been various studies that have overcome the issues and SVM currently has still become one of the algorithms that is widely adopted in solving various classification and prediction problems.

3. Methodology

3.1 Data Collection

This study has scrapped the Twitter data using the Twitter API Tweepy, from March to August 2022. The scrapped tweets are in English, using the keywords which are “COVID vaccine children”, “5 to 11 vaccine”, “Pfizer children” and “coronavac 5 to 11”. The extracted tweets contain noise such as stop words, emojis and URLs, which would be removed during pre-processing. The amount of collected tweets was 1500 and was stored in the csv format. Table 2 represents the example of a dataset that was extracted by Twitter API Tweepy.

Table 2
The example of raw data

Created at	Tweets	Username
2021-16-11 8:49:01	Disappointing to read that Health Canada approval of Pfizer vaccine for kids 5 to 11 may not come until December.	@IrfanDhalla
2022-12-1 10:28:12	Children didn’t seem to be getting COVID until they started pushing vaccines on them	@janet46195044
2022-12-1 10:19:08	Kids don’t need masks or vaccines. Stop torturing children!!	@evisokey

3.2 Data Pre-processing

As previously stated, the gathered tweets contain a few unwanted information, such as misspelt words, emojis, and abbreviated words. These words can interfere with the proper functioning of the model [17]. Therefore, it becomes crucial to remove all these unwanted words before moving to the next steps. The data pre-processing includes removing URL’s, stop words, and symbols, tokenization, case folding, handling non-standardized words, and stemming [25].

The unimportant characters are useless for text classification. In this step, the unimportant special characters, such as mentions (@username), retweets and hashtags (#) are removed. After that, comes the process of removal of the emojis, extra blank space, punctuations such as brackets and commas and also URLs, as these are irrelevant data. Table 3 shows the example of the tweets with the romoved noisy entity.

Case folding is the reducing all of the letters into small letters. This is to ensure that every word; regardless cases, are considered the same by the computer. For example, ‘COVID’ and ‘covid’ are considered the same word. Table 4 shows the example of tweets with the case folding. Stop words are also being removed in order to improve the performance of the classifier model. Stop words are the most used words in the English language such as articles (“an”, “a”, “the), verbs-to-be (“is”, “are”, “was”, “were”, “am”) and many more that does not represent any information. Stop word removal is removing all words that is likely to be irrelevant [5] and can save the processing time [1].

The tokenization is the process of segmenting text taken from a sentence or paragraph into certain parts [5]. It will break the sentence into single words, which is called tokens. It is performed to better understand the meaning of each sentence or tweet. Lemmatisation is a procedure that transforms words into the lemma, which is the root word. With this step, all the tokens are converted into their root word [26]. The “WordNet-Lemmatizer” is a useful function in the NLTK library of python that performs lemmatisation on the words given to it [19]. The lemmatisation is utilized for this project as lemmatisation considers the context when converting words into their root words, rather than just removing the last few letters as in stemming. Table 5 shows the example of the removed stop words and the tokenized words of the tweets.

Table 3
Removal of noisy entity

Original Tweets	Removal of Noisy Entity
vaccine for children is bad #stayawayfromourkids	vaccine for children is bad
Alhamdulillah my kid got their PFIZER shot today @fifiyzharuiddin7861	Alhamdulillah my kid got their PFIZER shot today
Kids don’t need masks or vaccines. Stop torturing children!!	Kids don’t need masks or vaccines stop torturing children
My children can finally go to school after vaccines	My children can finally go to school after vaccines

Table 4
Case folding

Original Tweets	Case Folding
vaccine for children is bad	vaccine for children is bad
Alhamdulillah my kid got their PFIZER shot today	alhamdulillah my kid got their pfizer shot today
Kids don’t need masks or vaccines stop torturing children	kids dont need masks or vaccines stop torturing children
My children can finally go to school after vaccines	my children can finally go to school after vaccines

Table 5
Stop words removal and tokenized words

	Original Tweets	Remove Stop Words	Tokenized Words
1.	vaccine for children is bad	vaccine children bad	[vaccine, children, bad]
2.	alhamdulillah my kid got their pfizer shot today	alhamdulillah kid got Pfizer shot today	[alhamdulillah, kid, got, pfizer, shot, today]
3.	kids don't need masks or vaccines stop torturing children	kids not need masks vaccines torturing children	[kids, not, need, masks, vaccines, torturing, children]
4.	my children can finally go to school after vaccines	children go to school vaccines	[children, go, school, vaccines]

3.3 Data Labelling

The supervised Machine Learning algorithm such as Support Vector Machine needs the data to be labelled into class. Therefore, the TextBlob and VADER, which is the existing library in Python have been utilized for this study. TextBlob would classify whether the word is positive or negative using predefined library. By using TextBlob, the polarity of a text will range between -1 and 1, where -1 represents negative sentiment while 1 represents positive sentiment respectively [4]. On the other hand, VADER works by finding percentage on how much positivity, negativity, and neutrality each sentence possessed. A compound score would be calculated, and final sentiment is determined. In this study, both tools have been tested in order to see which one performs better in the data labelling. According to [4], VADER could pick up the negative sentiment more than TextBlob. VADER is also more suitable in the labelling for social media data in sentiment analysis projects [27].

3.4 Feature Extraction

Machine learning algorithms are unable to use text as an input [26]. The Support Vector Machine is no exception as the algorithm can only analyze numerical data for classification. Therefore, the tweets need to be converted to features in order to enable them to be the input into the predictive modelling algorithms. The vectorization is performed by the n-grams model in which n consequent words are represented as one feature. A series of tokens, length n is defined as n-gram [28]. Unigram is taking one word as one feature at a time, whereas bigram will group two consequent words as one feature, and trigram is grouping three consequent words as one feature [26]. For this study, unigram is selected.

The feature extraction used in this study is the TF-IDF Vectorizer. The feature extraction uses TF-IDF method and generate term-document matrix, TF and IDF matrix that is useful during the classification stage. TF-IDF utilizes a weight to each which shows the importance of that word in a corpus. Therefore, TF-IDF consists of two metrics. The first metric is the term frequency that calculates the frequency of a word in a document. On the other hand, the second metric is the inverse document frequency that measures the frequency of documents that contain the word. Therefore, TF-IDF is obtained by multiplication of the TF and IDF. It can be explained by the following Eq. (1).

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D) \tag{1}$$

Term frequency of a particular term (t) is calculated as number of times a word occurs in a tweet. On the other hand, IDF is used to calculate the significance of the word, which carries meaningful context in classification. Based on Eq. (1), d represents the documents in the corpus D. Since the n-gram used is unigram (n=1), the words do not need to be transformed from the tokenization process earlier. Each importance of the word is calculated. The example of the weight values of the words after the TF-IDF calculation are as shown in Table 6. Based on the table, each importance of the word has been calculated and shown in the Rank column. In TF-IDF, the higher the number, the more important or significant the word is.

Table 6
Example Of TF-IDF Implementation

index	Terms	Rank
(5,0)	vaccine	0.00021
(1,0)	children	0.00004
(2,3)	torture	0.78985

3.5 System Architecture

The system architecture for this study is illustrated in Figure 2. The study begins with the data collection and data pre-processing. In the data collection, the data is scrapped from Twitter using the Tweepy. The data is then pre-processed in order to obtain the cleaned form. The next phase is the data labelling and VADER has been selected to identify the polarity into positive or negative data. After labelling, the feature extraction is done based on unigram model using the TF-IDF method. Then the training and testing of data are conducted using SVM algorithm based on the hold out method. There are 3 percentage of splits that have been selected for the training and testing of data. The best split is then chosen to be used for the SVM model development. In this study, linear kernel has been applied for the SVM model. The algorithm is then evaluated using the confusion matrix in order to obtain the classification performance before the model development. Once the SVM model is completed, the user could use the prototype for the sentiment classification via the graphical user interface. The result of the sentiment is divided into accepting and non-accepting, which represent the positive and negative sentiments respectively.

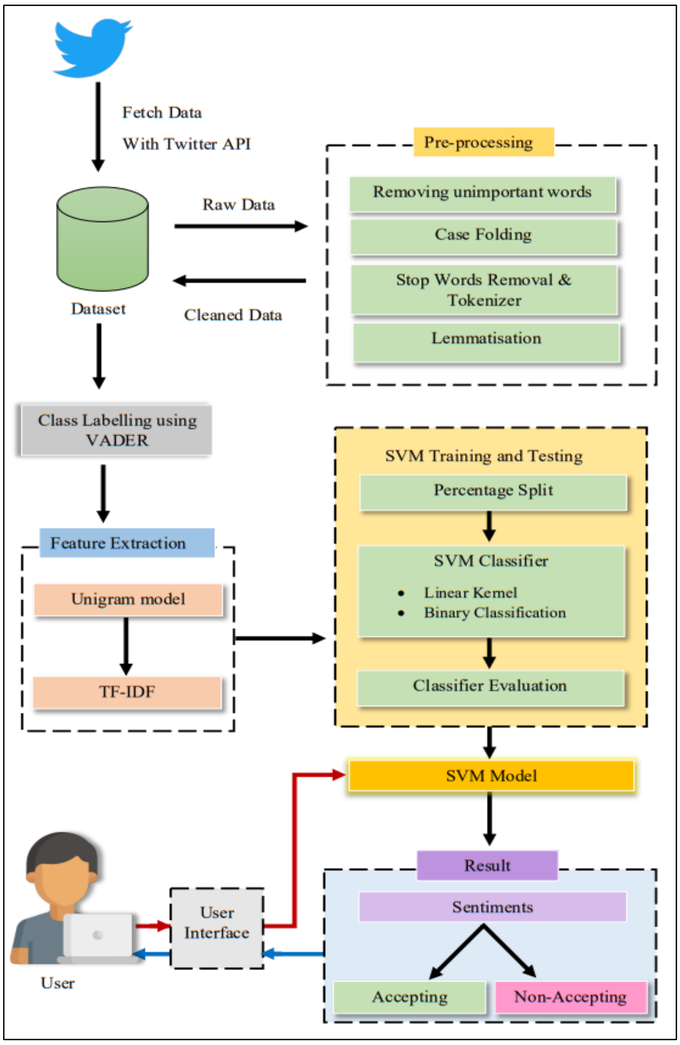


Fig. 2. System architecture of the project

3.6 Performance Evaluation

The performance evaluation has been based on the Confusion Matrix. The Confusion Matrix calculates the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP represents the correctly predicted positive sentiments as positive, whereas FP shows incorrectly predict negative sentiments as positive. TN represents the accurate predicted negative sentiments as negative, while FN represents the incorrectly positive sentiments as negative. Based on these metrics, the results of accuracy, precision, sensitivity (or recall), and F-score could be obtained.

The recall is also known as sensitivity and true positive rate (TPR). The value is preferable when it is higher [29]. The formula for recall is as shown in Eq. (2). It represents the percentage or proportion of correctly classified positive sentiments among all actual positive sentiments. Specificity is also known as true negative rate. The formula for specificity is as shown in Eq. (3). It represents the percentage or proportion of actual negative sentiments among all classified negative sentiments.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (2)$$

$$\text{Specificity} = \frac{TN}{N} \qquad (3)$$

The formula for precision is as shown in Eq. (4). It represents the percentage or proportion of how many actual positive sentiments among all classified positive sentiments. The formula for accuracy is as shown in Eq. (5). Accuracy shows the percentage or proportion of correctly classified class.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

$$\text{Accuracy} = \frac{TP+TN}{P+N} \tag{5}$$

The false positive rate (FPR) is the complimentary to recall. The formula for false positive rate is shown in Eq. (6). False positive rate shows the percentage of negative cases incorrectly identified as positive cases in the data. The F-score is also known as F-Measures, as shown in Eq. (7). It is the value that represents the mean between the precision and recall.

$$\text{False Positive Rate} = \frac{FP}{FN+TN} \tag{6}$$

$$\text{F-score} = \beta\,\frac{Precision\ x\ Recall}{Precision+Recall} \tag{7}$$

The values in the confusion matrix are also presented in the Receiver Operating Characteristic (ROC) curve. It is a plot that evaluates the performance of binary classification model on positive class [30]. It is a graph of sensitivity against the precision. The area above the graph is abbreviated as Area Under the Curve (AUC), which determines the performance of the model. The higher the value of AUC, the better the model in classifying the negative and positive sentiments in the sentiment analysis. The further the curve from the 45-degree diagonal of the ROC space, the more accurate the model [31].

4. Results and Discussion

There are two analyses that have been conducted in this study. The first analysis conducted was the exploratory data analysis on the collected tweets related to the COVID-19 vaccination for children. The second analysis is on the performance of the SVM classifier.

4.1 Explanatory Data Analysis

The explotary data analysis includes the analysis from the data labelling and the word cloud. In this study, the data labelling using VADER has produced better results than using TextBlob. The data has been labelled to positive and negative, which represent the 2-class classification.

Figure 3 shows the bar chart of comparison between the number of Positive and Negative data labelling using VADER. The Positive data represents the positve tweets on children's COVID-19 vaccination, while the negative represents the opposite. Based on the figure, it could be seen that there are more positive tweets about the children's vaccination than the negative ones. This shows that actually more people are positive in giving COVID-19 vaccine to their children.

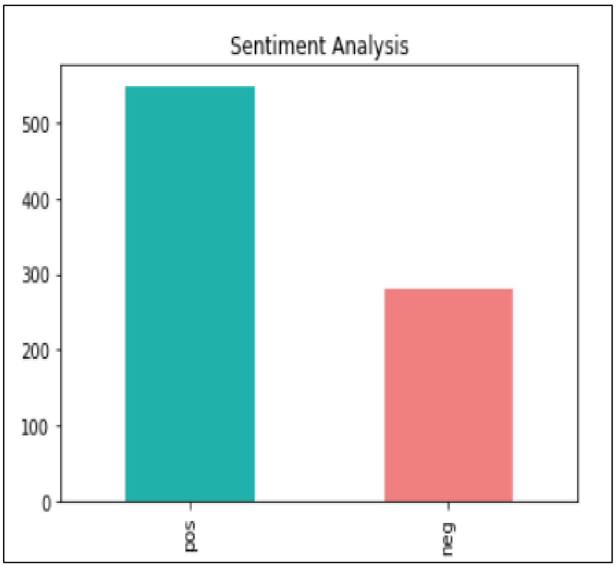
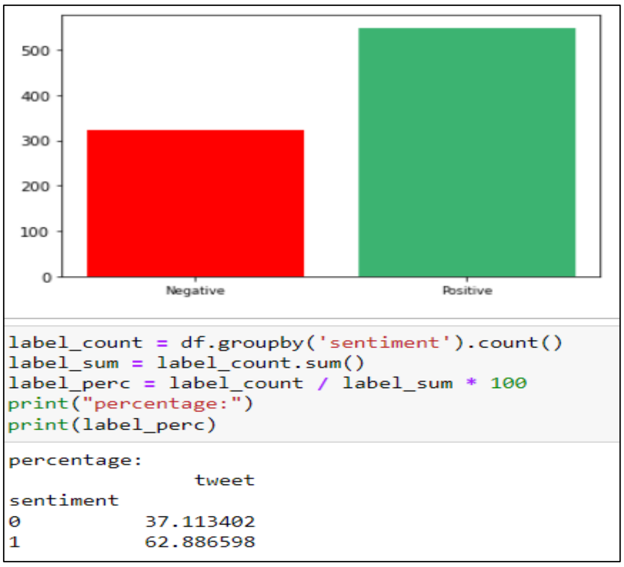


Fig. 3. Data labelling using VADER

In data labelling, the class balance may possibly be improved by resampling or using library to oversample the negative sentiment. In training and testing any machine learning model, the class should be balanced or not too different from one another. However, it is common for a sentiment data to have imbalance data especially from scraping [32]. In this study, the positive data was almost twice the negative data. In order to solve this problem, there were two methods used in this project, which were the oversampling using Synthetic Minority Oversampling Technique (SMOTE), and resampling. In SMOTE, the minority data is duplicated to produce new data of the smaller number of class. In resampling, more data is scrapped from the twitter as the additional dataset.

Figure 4 shows the data labelling after using SMOTE. Based on Figure 4, the number of negative data has been increased and improved compared to the data labelling as shown in Figure 3. After using SMOTE, the new percentage of positive data is 62.89%, while the negative data is 37.11%. The number of overall data is 873, which consisted of 549 positive and 324 negative class



```
label_count = df.groupby('sentiment').count()
label_sum = label_count.sum()
label_perc = label_count / label_sum * 100
print("percentage:")
print(label_perc)
```

```
percentage:
                tweet
sentiment
0           37.113402
1           62.886598
```

Fig. 4. Data labelling after using SMOTE

After the data labelling is completed, the word clouds are generated to represent the most frequent words appearing in each class. Figure 5 shows the distribution of the words in the negative class. The bigger words mean the more frequent words. Based on the word cloud, the most frequently mentioned words are “pfizer”, “vaccine” and “kid”. There are words such as “death” and “kill” which clearly represent negative sentiments. Figure 6 shows the word cloud for the positive class. Based on the word cloud, there are words that indicate positive sentiment, such as “safe” and “need”. The frequently used words in the positive cloud also include “pfizer”, “covid”, and “kid”. These overlapping words are actually neutral and would need another connecting word for the classification of positive or negative class.

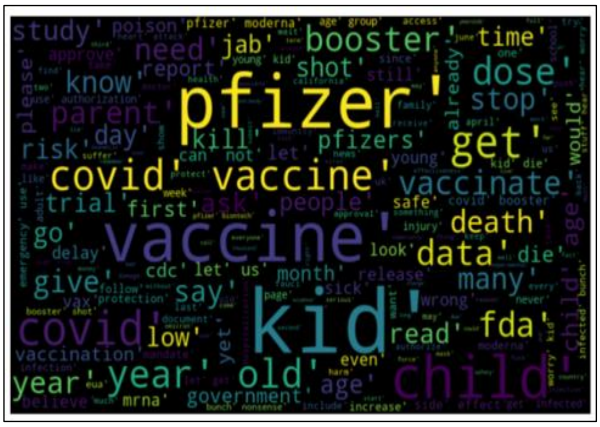


Fig. 5. The word cloud for negative class

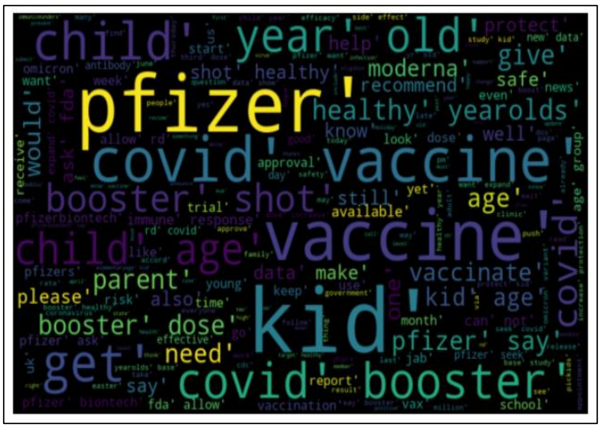


Fig. 6. The word cloud for positive class

4.2 Support Vector Machine Performance Evaluation

In this study, the holdout method with the three percentage splits have been experimented with the dataset. Aside from the percentage split, the result from the resampled data and oversampled dataset have also been recorded. The experiments have been based on three set of dataset arragement. The original dataset is called the Dataset 1, the resampled dataset is called Dataset 2 and finally the dataset that applies SMOTE is called the Dataset 3. Another evaluation is the AUC of the ROC, which could identify the model’s ability to distinguish between the positive class and the negative class.

The True Positive, True Negative, False Positive and False Negative values have been obtained from the python library sklearn. Figure 7 shows the Confusion Matrix for the model when trained and tested with 80:20 data split using Dataset 1. Figure 8 shows the calculation of Accuracy, Precision, Recall, F1 Score, Specificity and False Positive Rate by using values in Figure 7. While all the percentage values are higher, the specificity is significantly lower. The low specificity means that the model miss predicts many negative sentiments test data as positive. This low percentage may be a result from the lower number of negative-class data in Dataset 1.

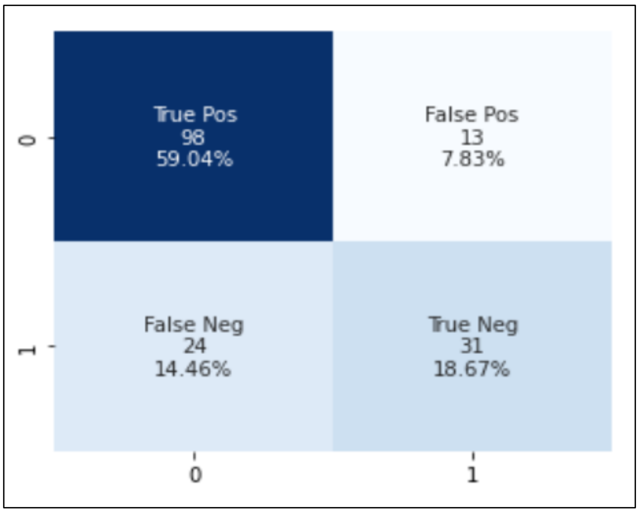
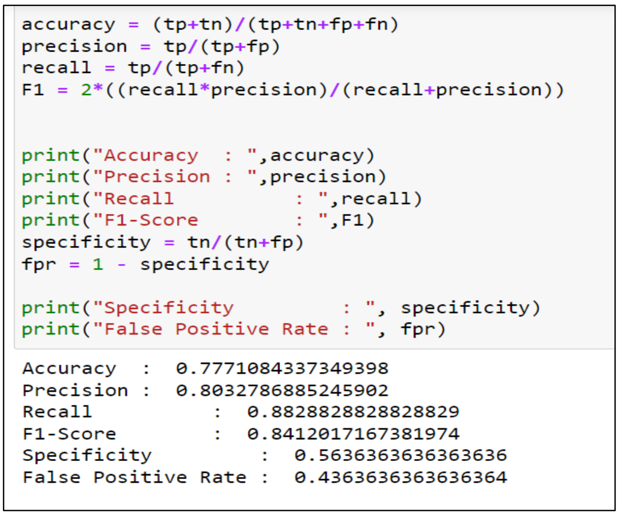


Fig. 7. Confusion matrix for dataset 1 with 80:20 split



```
accuracy = (tp+tn)/(tp+tn+fp+fn)
precision = tp/(tp+fp)
recall = tp/(tp+fn)
F1 = 2*((recall*precision)/(recall+precision))


print("Accuracy  : ",accuracy)
print("Precision : ",precision)
print("Recall          : ",recall)
print("F1-Score        : ",F1)
specificity = tn/(tn+fp)
fpr = 1 - specificity

print("Specificity         : ", specificity)
print("False Positive Rate : ", fpr)
```

```
Accuracy  :  0.7771084337349398
Precision :  0.8032786885245902
Recall          :  0.8828828828828829
F1-Score        :  0.8412017167381974
Specificity         :  0.5636363636363636
False Positive Rate :  0.4363636363636364
```

Fig. 8. Calculation for accuracy of dataset 1 with 80:20 split

Table 7 shows the evaluation results of the three different datasets based on the three different percentage splits, 80:20, 90:10 and 70:30. Based on the table, the highest accuracy is obtained from the Dataset 2 when trained and tested with the 70:30 split, and Dataset 3 with the 90:10 split. Although the accuracies are the same (82%), their AUC results are different. The model that has been trained and tested using Dataset 3 with 90:10 split has outperformed the one with the Dataset 2 (70:30 split). Nevertheless, the precision, recall, F1-score, specificity, and false positive rate have been calculated to investigate the best model between the two for the study.

Table 7
The evaluation results of three different splits

	70:30		80:20		90:10	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Dataset 1	76%	0.718	78%	0.723	78%	0.718
Dataset 2	82%	0.777	79%	0.777	82%	0.815
Dataset 3	79%	0.766	79%	0.777	82%	0.810

Based on the Table 8, the Dataset 3, which is dataset with the SMOTE implementation, has produced better performance overall when trained and tested using the 90:10 percentage split. It should be noted that false positive rate should be smaller, and not bigger. Therefore, the performance of the model is good and acceptable, with its high precision and recall. This shows that the positive and negative classes are well classified by the model.

Table 8
The comparison of two models

	Precision	Recall	Specificity	F1 Score	False Positive Rate
Dataset 2 (70:30)	0.841	0.851	0.682	0.846	0.318
Dataset 3 (90:10)	0.891	0.830	0.793	0.860	0.206

Figure 9 shows the confusion matrix for the best accuracy model, which is the Dataset 3 with 90:10 percentage split. Based on the confusion matrix, the model correctly classifies 49 tweets as positive, falsely classifies 10 tweets as positive, correctly classifies 23 tweets as negative, and falsely classifies 6 as negative.

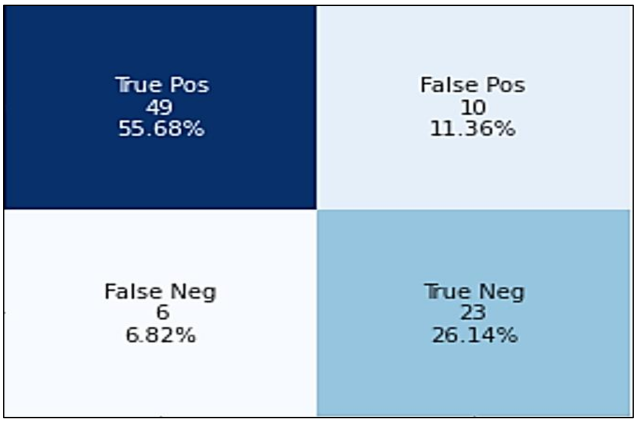


Fig. 9. Confusion matrix of dataset 3 with 90:10 split

Figure 10 shows the ROC curve and AUC also for the best model accuracy. The AUC for ROC curve is 0.8118. The further the curve from the 45-degree diagonal (red dotted line) of the ROC space, the more accurate the model. The higher the value of AUC, the better the model is in classifying the negative and positive sentiments in Sentiment Analysis. The results show that the SVM model could produce good performance in the sentiment classification.

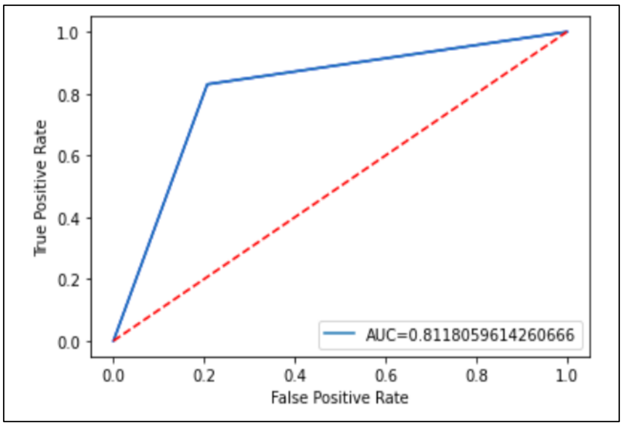
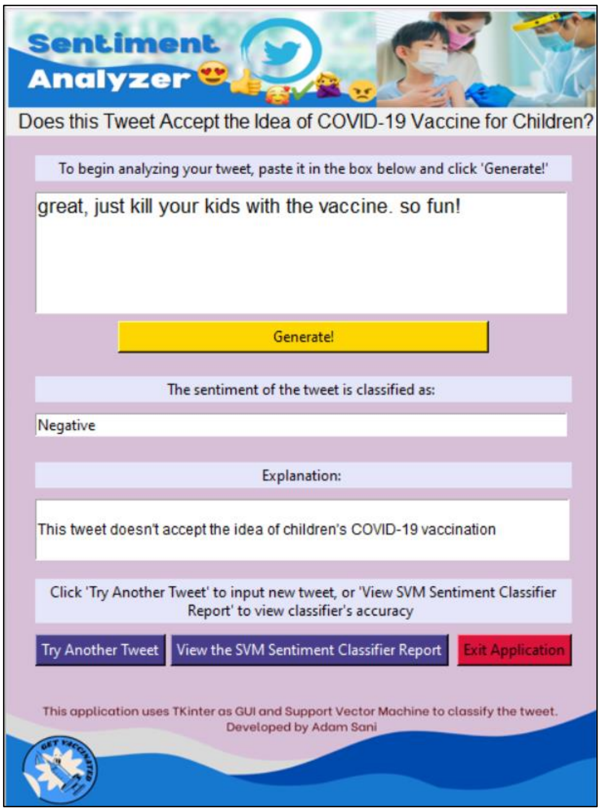


Fig. 10. ROC and AUC of dataset 3 with 90:10 split

In this study, the best accuracy achieved is 82% and this is aligned with other SVM performance in COVID-19 based sentiment classifications. Other studies have also produced accuracies of more than 80% in the sentiment classifications [9][10][25][26]. It could be concluded that the level of accuracy, precision, recall and specificity of the SVM model in this study are sufficient enough to develop a good classifier for this specific problem.

4.3 Proposed Prototype and Evaluation

In order to build a useful application for the sentiment analysis, a GUI has been developed using the Tkinter to test the SVM model. Figure 11 shows the screenshot of the GUI when the negative sentiment tweet is entered. The user must click on the button “Generate!” to view the result of the sentiment of the tweet entered. There is an explanation text field to further clarify the negative sentiment. The “Try Another Tweet” button will clear all the text fields and then the input section will be ready to receive a new tweet. The exit button is self-explanatory, which will close the window of the application. The “View the SVM Sentiment Classifier Report” will open a new window where the performance of the classifier will be shown.



Does this Tweet Accept the Idea of COVID-19 Vaccine for Children?

To begin analyzing your tweet, paste it in the box below and click 'Generate!'

great, just kill your kids with the vaccine. so fun!

Generate!

The sentiment of the tweet is classified as:

Negative

Explanation:

This tweet doesn't accept the idea of children's COVID-19 vaccination

Click 'Try Another Tweet' to input new tweet, or 'View SVM Sentiment Classifier Report' to view classifier's accuracy

Try Another Tweet | View the SVM Sentiment Classifier Report | Exit Application

This application uses TKinter as GUI and Support Vector Machine to classify the tweet.
Developed by Adam Sani

Fig. 11. Screenshot of the system prototype

To further investigate the model's performance in the sentiment classification, three more tweets have been chosen from the Twitter as the input texts. One of them possesses obvious positive sentiment, one possesses obvious negative sentiment and one of them are chosen from sarcastic negative tweet. Table 9 shows the comparison of the actual sentiment and the classification output of the SVM model. The model has successfully classified the tweets into the correct class even there are sarcasm words in them. The successful classification has proven that the model is good at classifying tweets related to the acceptance of COVID-19 vaccine for children.

Table 9
The result of each test Input

	Input Scenarios	Actual Sentiment	Classification Output
1.	Great, just kill your kid with the vaccine. So fun!	Negative	Negative
2.	The vaccine is safe as it already recommended by the FDA. I don't know what's the problem.	Positive	Positive
3.	I personally think that the Pfizer is dangerous for vulnerable child.	Negative	Negative

4.4 Research Limitation

The first limitation is that the model could only classify English tweets. In this study, only English tweets are scraped and thus, the SVM model is trained and tested based on the English-language sentences dataset. It is considered as a limitation since the conversations about COVID-19 for children are not only discussed in English language, but also in other languages around the world. However, there is also a disadvantage if the tweets are non-English. The disadvantage is that it will add to the complexity of data preparation stage as the stop words may be not supported by the NLTK corpus. The amount of scrapped data for non-English tweets could also be small and limited to be analyzed with the machine learning.

The next limitation is the number of scraped data in this study is quite small. This limitation is due to the time constraint of the period of the study and the limitation of the specific topic of children vaccination being discussed over Twitter. It is expected that if larger amount of data are used, the higher the accuracy of the model will be. However, in this study, the SVM model could also produced acceptable results with the amount of data collected. This shows that the amount of the dataset is good enough for SVM to achieve good results for this sentiment classification.

5. Conclusions

This study has explored the capability of SVM in the sentiment classification of the COVID-19 children vaccination acceptance based on the Twitter data. Based on the experimental results, the SVM model has generated acceptable performance in this classification problem. The SVM model has successfully classified the positive and negative tweets of the children vaccination with good accuracy. This study has contributed to another finding of the capability of SVM in solving the sentiment classification problems. SVM has proven to be able to produce good performance in different sentiment classification problems.

The significance of the study is that the results could be used by the public and authorities to determine the public's view on the COVID-19 children vaccination. The authority that may benefit from this study is the government and health agencies. This is specially to investigate the public sentiments about the acceptance of the COVID-19 vaccine for children. It can help the government to decide whether to increase awareness on the vaccination programme. The parents may also benefit from this study as they could be aware with what other people's opinions regarding the implementation of COVID-19 vaccination for children. In a way, it could increase the awareness of people to get their children the COVID-19 vaccine shots. Based on the Twitter data in this study, it is found that actually more people are positve about the COVID-19 vaccination on children.

The future work would be collecting more data from different sources of social media such as the Facebook and Reddit. This could extend the study to larger demographic view for better representation of the sentiments. Another work is to compare SVM with other machine learning techinques in order to further investigate the algorithm's capability and performance.

Acknowledgement
The authors would like to thank Universiti Teknologi MARA Cawangan Terengganu for the continuous support given to this research. This research is funded under the UiTM Geran Penyelidikan Myra Lepasan PHD (600-RMC/GPM LPHD 5/3 (068/2021).

References

[1] Alharbi, Najla Hamandi, and Jawad Hassan Alkhateeb. "Sentiment analysis of arabic tweets related to covid-19 using deep neural network." In *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, pp. 1-11. IEEE, 2021. https://doi.org/10.1109/ICOTEN52080.2021.9493467

[2] Vanaja, Satuluri, and Meena Belwal. "Aspect-level sentiment analysis on e-commerce data." In *2018 International conference on inventive research in computing applications (ICIRCA)*, pp. 1275-1279. IEEE, 2018. https://doi.org/10.1109/ICIRCA.2018.8597286

[3] Iksan, Nur, Djoko Adi Widodo, Budi Sunarko, Erika Devi Udayanti, and Etika Kartikadharma. "Sentiment analysis of public reaction to COVID19 in twitter media using naïve Bayes classifier." In *2021 IEEE International conference on health, instrumentation & measurement, and natural sciences (InHeNce)*, pp. 1-4. IEEE, 2021. https://doi.org/10.1109/InHeNce52833.2021.9537243

[4] Rahul, Kumar, Bhanu Raj Jindal, Kulvinder Singh, and Priyanka Meel. "Analysing public sentiments regarding COVID-19 vaccine on twitter." In *2021 7th international conference on advanced computing and communication systems (ICACCS)*, vol. 1, pp. 488-493. IEEE, 2021. https://doi.org/10.1109/ICACCS51430.2021.9441693

[5] Nurdeni, Deden Ade, Indra Budi, and Aris Budi Santoso. "Sentiment analysis on Covid19 vaccines in Indonesia: from the perspective of Sinovac and Pfizer." In *2021 3rd East Indonesia conference on computer and information technology (EIConCIT)*, pp. 122-127. IEEE, 2021. https://doi.org/10.1109/EIConCIT50028.2021.9431852

[6] Liu, Siru, and Jialin Liu. "Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis." *Vaccine* 39, no. 39 (2021): 5499-5505. https://doi.org/10.1016/j.vaccine.2021.08.058

[7] Kidam, Kamarizan, Siti Aishah Rashid, Jafri Mohd Rohani, Hafizah Mahmud, Hamidah Kamarden, Fateha Abdul Razak, Nurul Nasuha Mohd Nor, and Nur Kamilah Abdul Jalil. "Development of Instrument to Measure the Impact of COVID-19 And Movement Control Order to Safety and Health Competent Person and Training Provider." *Journal of Advanced Research in Technology and Innovation Management* 2, no. 1 (2022): 22-28.

[8] Gopi, Rahul Sanmugam, Lavanya Dhanesh, Mohammad Aljanabi, Tavanam Venkata Rao, M. Thiruveni, and S. Mahalakshmi. "Design of Covid19 disease detection for risk identification using deep learning approach." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 32, no. 1 (2023): 139-154. https://doi.org/10.37934/araset.32.1.139154

[9] Berkeley, L. J. "*FDA authorizes Pfizer's Covid vaccine for kids ages 5 to 11, shots could begin early next week with CDC clearance.* " *CNBC Health and Science.* (2021)

[10] LaFraniere, S. and Weiland, N. " *F.D.A. Panel Recommends Covid Shots for Children 5 to 11.* "*New York Times.* (2021)

[11] Neel, J., Wroth, C., and Greenhalgh, J. "*CDC recommends Pfizer's COVID vaccine for children ages 5 through 11.* " *National Public Radio.* (2021)

[12] Sick-Samuels, A. C., and Messina, A. "*COVID Vaccine: What Parents Need to Know.* " *John Hopkins Medicine.*(2021)

[13] *JCVI statement on COVID-19 vaccination of children aged 12 to 15 years: 3 September 2021*. (2021). *Department of Health & Social Care UK*.

[14] Tank, S. *Pfizer says Covid vaccine more than 90% effective in kids.CNBC Health and Science.* (2021)

[15] Yousefinaghani, Samira, Rozita Dara, Samira Mubareka, Andrew Papadopoulos, and Shayan Sharif. "An analysis of COVID-19 vaccine sentiments and opinions on Twitter." *International Journal of Infectious Diseases* 108 (2021): 256-262. https://doi.org/10.1016/j.ijid.2021.05.059

[16] Prakash, T. Nikil, and A. Aloysius. "A Comparative study of Lexicon based and Machine learning based classifications in Sentiment analysis." *International Journal of Data Mining Techniques and Applications* 8, no. 1 (2019): 43-47. https://doi.org/10.20894/IJDMTA.102.008.001.005

[17] Banik, Sagar, Aniket Ghosh, Sumit Banik, and Anupam Mukherjee. "Classification of COVID19 Tweets based on sentimental analysis." In *2021 International conference on computer communication and informatics (ICCCI)*, pp. 1-7. IEEE, 2021. https://doi.org/10.1109/ICCCI50826.2021.9402540

[18] Ezhilan, Aakash, R. Dheekksha, R. Anahitaa, and R. Shivani. "Sentiment analysis and classification of COVID-19 tweets." In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 821-828. IEEE, 2021. https://doi.org/10.1109/ICOEI51242.2021.9453062

[19] Jayasurya, Gutti Gowri, Sanjay Kumar, Binod Kumar Singh, and Vinay Kumar. "Analysis of public sentiment on COVID-19 vaccination using twitter." *IEEE Transactions on Computational Social Systems* 9, no. 4 (2021): 1101-1111. https://doi.org/10.1109/TCSS.2021.3122439

[20] Nezhad, Zahra Bokaee, and Mohammad Ali Deihimi. "Twitter sentiment analysis from Iran about COVID 19 vaccine." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 16, no. 1 (2022): 102367. https://doi.org/10.1016/j.dsx.2021.102367

[21] Mohan, Lalit, Janmejay Pant, Priyanka Suyal, and Arvind Kumar. "Support vector machine accuracy improvement with classification." In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 477-481. IEEE, 2020. https://doi.org/10.1109/CICN49253.2020.9242572

[22] Sharma, Shashank, Sumit Srivastava, Ashish Kumar, and Abhilasha Dangi. "Multi-class sentiment analysis comparison using support vector machine (svm) and bagging technique-an ensemble method." In *2018 International conference on smart computing and electronic enterprise (ICSCEE)*, pp. 1-6. IEEE, 2018. https://doi.org/10.1109/ICSCEE.2018.8538397

[23] Prastyo, Pulung Hendro, Igi Ardiyanto, and Risanuri Hidayat. "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF." In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1-6. IEEE, 2020. https://doi.org/10.1109/ICDABI51230.2020.9325685

[24] Goyal, C. *Multiclass Classification Using SVM.Analytics Vidhya*. (2018)

[25] Iksan, Nur, Djoko Adi Widodo, Budi Sunarko, Erika Devi Udayanti, and Etika Kartikadharma. "Sentiment analysis of public reaction to COVID19 in twitter media using naïve Bayes classifier." In *2021 IEEE International conference on*

health, instrumentation & measurement, and natural sciences (InHeNce), pp. 1-4. IEEE, 2021. https://doi.org/10.1109/InHeNce52833.2021.9537243

[26] Ghasiya, Piyush, and Koji Okamura. "Investigating COVID-19 news across four nations: A topic modeling and sentiment analysis approach." *Ieee Access* 9 (2021): 36645-36656. https://doi.org/10.1109/ACCESS.2021.3062875

[27] Bakharia, A. *Quick Social Media Sentiment Analysis with VADER. Medium.* (2016)

[28] Rahman, Sheikh Shah Mohammad Motiur, Khalid Been Md Badruzzaman Biplob, Md Habibur Rahman, Kaushik Sarker, and Takia Islam. "An investigation and evaluation of N-Gram, TF-IDF and ensemble methods in sentiment classification." In *Cyber Security and Computer Science: Second EAI International Conference, ICONCS 2020, Dhaka, Bangladesh, February 15-16, 2020, Proceedings 2*, pp. 391-402. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-52856-0_31

[29] Nerkhede, S. *Understanding Confusion Matrix. Towards Data Science.* (2018)

[30] Bhandari, A. *AUC-ROC Curve in Machine Learning Clearly Explained. Analytics Vidhya.* (2020)

[31] Chan, C. *What is a ROC Curve and How to Interpret It. Displayr.*(2018)

[32] Jiang, H. Sentiment Analysis on Imbalanced Airline Data. (2016) *https://hmjianggatech.github.io/files/BHAMProject/SentimentAnalysis.pdf*