information

Article

Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes

Charlyn Villavicencio [1,2,*], Julio Jerison Macrohon [1], X. Alphonse Inbaraj [1], Jyh-Horng Jeng [1] and Jer-Guang Hsieh [3]

[1] Department of Information Engineering, I-Shou University, Kaohsiung City 84001, Taiwan; isu10903050D@cloud.isu.edu.tw (J.J.M.); xalphonse@gmail.com (X.A.I.); jjeng@isu.edu.tw (J.-H.J.)
[2] College of Information and Communications Technology, Bulacan State University, Bulacan 3000, Philippines
[3] Department of Electrical Engineering, I-Shou University, Kaohsiung City 84001, Taiwan; jghsieh@gmail.com
* Correspondence: charlyn.villavicencio@bulsu.edu.ph; Tel.: +886-958-450-028

Abstract: A year into the COVID-19 pandemic and one of the longest recorded lockdowns in the world, the Philippines received its first delivery of COVID-19 vaccines on 1 March 2021 through WHO's COVAX initiative. A month into inoculation of all frontline health professionals and other priority groups, the authors of this study gathered data on the sentiment of Filipinos regarding the Philippine government's efforts using the social networking site Twitter. Natural language processing techniques were applied to understand the general sentiment, which can help the government in analyzing their response. The sentiments were annotated and trained using the Naïve Bayes model to classify English and Filipino language tweets into positive, neutral, and negative polarities through the RapidMiner data science software. The results yielded an 81.77% accuracy, which outweighs the accuracy of recent sentiment analysis studies using Twitter data from the Philippines.

Keywords: COVID-19; COVID-19 vaccine; Naïve Bayes; natural language processing; sentiment analysis; twitter; tweets

Citation: Villavicencio, C.; Macrohon, J.J.; Inbaraj, X.A.; Jeng, J.-H.; Hsieh, J.-G. Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes. *Information* **2021**, *12*, 204. https://doi.org/10.3390/info12050204

Academic Editor: Arkaitz Zubiaga

Received: 18 April 2021
Accepted: 7 May 2021
Published: 11 May 2021

1. Introduction

COVID-19 has greatly affected several areas in our lives, the environment, profit, mental health, and public transportation. In the year 2020, the global economy contracted by 3% with a significant loss of around USD 9 trillion. Only 34% of enrolled students in the world were able to enjoy the benefits of physical or proper education, and various business sectors experienced a great decline in productivity and employment [1]. The paper titled "The Mental Health Consequences of COVID-19 and Physical Distancing" stated that COVID-19 has brought short- and long-term consequences in people's mental health and wellbeing. The authors also stated that just as the severe acute respiratory syndrome (SARS) was associated with patients and medical personnel having post-traumatic stress disorder (PTSD), stress, and psychological distress, COVID-19 appears to have triggered anxiety, depression, loneliness, violence, and substance abuse. Furthermore, because the schools have been closed, there is a real possibility of child abuse [2].

COVID-19 has also resulted in a significant decline in public transportation demand and revenue, which has not happened before, because of government orders and personal choices as people refrained from traveling to minimize the risk of acquiring the virus. Thus, there has been a dramatic impact on lifestyle and travel worldwide, ranging from the decline of air travel services to the increase in demand for online communication [3]. India is suffering from a huge increase in positive COVID-19 cases, and there is fear of mass destruction and casualties. Because the virus is contagious, isolation of the infected person from the family and fear of losing family members are the consequences of COVID-19 [4].

A huge area of our lives has been affected by the COVID-19 pandemic. Fortunately, there are now vaccines available to provide immunity against the harm of the virus. In the

ASEAN region, the Philippines was the last to receive its COVID-19 vaccine shipment [5]. A total of 600,000 doses of China's Sinovac, which came from WHO's COVAX initiative and a donation from the People's Republic of China, arrived in the Philippines on 28 February 2021 [6]. A total of 525,600 doses of the Oxford–AstraZeneca vaccine also arrived 12 days later [7], with 400,000 more doses of China's Sinovac arriving 12 days after that [8].

On 1 March 2021, almost a year into the pandemic, the Philippines began its vaccination program, starting with health professionals [9]. The Philippine Department of Health devised its vaccine rollout plan titled "The Philippine National Deployment and Vaccination Plan for COVID-19 Vaccines" [10], also known as "ResBakuna". The document highlights the priority list of those who are eligible to receive the vaccine and is listed in Table 1.

Table 1. Priority eligible population group in the Philippines [10]. This table shows three groups (A, B, and C), which specifies the level of prioritization in the vaccination rollout of the Philippine government.

Priority	Population Group
Priority Group A	1. Frontline workers in health facilities. 2. Senior citizens aged 60 years and above. 3. Persons with comorbidities. 4. Frontline personnel in essential sectors.
Priority Group B	1. Teachers and social workers. 2. Other government workers. 3. Other essential workers. 4. Sociodemographic groups at significantly higher risk. 5. Overseas Filipino workers. 6. Other remaining workforce.
Priority Group C	Rest of the Filipino population not otherwise included in Priority Groups A and B.

Because the government devised a vaccination rollout plan, Filipinos are enjoined to be vaccinated to acquire immunity against the virus. However, individuals can freely choose whether or not to be vaccinated. In this context, the goal of this study was to analyze the sentiment of Filipinos towards COVID-19 vaccines through the social networking site Twitter and classify them into positive, neutral, and negative polarities. The results of this study can help the Philippine government to make wise decisions about fund allocation, vaccination provision, and strategic scheduling of its vaccination rollout plans. The proposed method can be applied to English and Tagalog tweets to classify them according to their polarity, which can be used for similar studies. Previous studies successfully utilized data from Twitter regarding the local airlines and COVID-19 in the Philippines. However, because the COVID-19 vaccination only came recently, there have been no studies regarding this issue. In line with this, the researchers made use of all the tweets in the first month of the implementation of the vaccination program. The main contribution of this study can be summarized as follows:

- It automatically labels the polarity of both English and Filipino language tweets.
- It reports the sentiments of Filipinos towards COVID-19 vaccines.
- The government can use this study as a tool to make wise decisions regarding the vaccination program.
- The proposed model can continuously analyze incoming tweets to monitor any updates or changes in the attitudes of Filipinos towards COVID-19 vaccines.

2. Related Literature

Nowadays, researchers are using posts on social media for analysis in order to achieve or predict results. Social media is a great platform to express sentiments, views, and opinions [11]. Twitter is one of the most widely used social media platforms in the world,

where users can post anything in their mind. Twitter has over 100 million active users [12], and the number of tweets posted every day can reach up to 500 million [13]. Twitter allows people to genuinely express themselves in a timely manner. This is different from traditional face-to-face interviews, where the interviewee's response may be affected because of the nervousness brought on by the live communication between the interviewer and interviewee [14]. When users are on their solitude, they can easily express themselves in a genuine manner that's why twitter is a good platform to use for analyzing true public sentiment [14]. Because users can freely share their location, comments, opinions, and feelings, albeit limited within 280 characters, it is suitable in studies that require opinion analysis. Moreover, due to Twitter's application programming interface (API) and database access being available to the public, the data collection can be easier [14].

Natural language processing (NLP) is used to retrieve information from a given text [11]. This is a process where the computer extracts meaning from sentences made by a human. NLP can be used in text mining, language translations, and programmed question answering [15], such as chat bots used by businesses to cater simultaneous customer queries. Sentiment analysis, which is also known as opinion mining, is a computational study of opinions, sentiments, and emotions conveyed in given words or sentences [16]. The researchers made use of preprocessed tweets using NLP techniques and classified the sentiment expressed by the tweets into positive, neutral, and negative polarities.

Different frameworks for sentiment analysis using Twitter data have been proposed. One of the methods is the attention-based bidirectional CNN–RNN deep model (ABCDM). This framework utilizes two independent bidirectional long short-term memory (LSTM) and gated recurrent unit (GRU) layers to extract past and future contexts by taking into consideration temporal information flow in both directions, which achieves state-of-the-art results for both short and long reviews [17].

Another method is the bidirectional emotional recurrent unit (BiERU) for conversational sentiment analysis, which extracts the sentiment of each message in a text conversation. This method proposes a fast, compressed framework based on emotional recurrent units with fewer parameters. The BiERU model is party-independent and thus suitable to be integrated in multiparty conversations without the need for adjustments [18].

Sentiment analysis is also useful in accurately detecting traffic accidents by utilizing social media posts and NLP techniques, as shown in the study by Ali et al. titled "Traffic Accident Detection and Condition Analysis Based on Social Networking Data". The authors of the study used ontology and latent Dirichlet allocation (OLDA) for topic labeling to extract traffic-related posts and discard other topics. They also trained bidirectional long short-term memory (Bi-LSTM) with SoftMax regression to classify texts according to their polarity, with all the necessary reports organized to be sent to the police station and emergency management office for immediate action [19].

In the medical field, sentiment analysis is also utilized to analyze posts relating to drug reviews or side effects on social media together with the patient's medical records to recommend personal diabetes and blood pressure (BP) healthcare treatment. The study by Ali et al. titled "An Intelligent Healthcare Monitoring Framework Using Wearable Sensors and Social Networking Data" used Bi-LSTM with ontologies to classify diabetes, BP, mental health, and side effects of medicine as well as Hadoop MapReduce with machine learning to reduce the size of data about patient treatments [20]. This healthcare monitoring framework promotes timely monitoring of diabetes and BP patients regarding their health condition before it worsens.

RapidMiner (RM) is a data science software consisting of data preprocessing techniques, machine learning algorithms, and model building operators [21]. A lot of NLP techniques are available in RM, such as case transformation, tokenization, stemming, stop words removal, etc., which preprocesses texts to obtain meaningful relationships between words and determine what the sentence implies.

Naïve Bayes is commonly used to solve classification issues. It has also been noted that Naïve Bayes performs accurately in determining the true polarity of a given sentence,

even in unbalanced datasets [22]. Moreover, this model is a high-bias and low-variance classifier that works well even in a small dataset [21]. Naïve Bayes comes from two words. Naïve comes from this method assuming that one occurrence of a certain feature is independent of the occurrence of other features. Thus, each feature contributes individually to classification without dependence on other features. Bayes comes from the principles of the Bayes' theorem [23] and this classifier calculates the probability of an event in a series of steps which will be discussed in Section 3.4. of this paper [24].

A previous study used Twitter data from the Philippines to examine local airline sentiments [15]. This study successfully determined the attitude of Filipinos towards the country's local airlines by applying NLP techniques and comparing three classifier algorithms. Samonte et al. used Naïve Bayes, support vector machine, and random forest to develop a model, and the results showed that Naïve Bayes yielded the highest accuracy (66.67%) in determining the true polarity of tweets. Inspired by this study, Abisado et al. [22] conducted research on Twitter sentiments of Filipinos during the COVID-19 pandemic using multinomial Naïve Bayes classifier, which revealed that 52% of Filipinos have a positive attitude and 48% have a negative attitude towards the pandemic; the classifier model yielded 72.17% accuracy.

Several research attempts have been made to determine the polarity of given texts through sentiment analysis, especially regarding the COVID-19 pandemic. This study focused on determining the stance of Filipinos regarding vaccination. Moreover, a classifier model was developed using the Naïve Bayes classification algorithm to classify the sentiments expressed in tweets relating to COVID-19 vaccines into positive, neutral, and negative polarities. The classifier analyzed tweets written in both English and Tagalog, which are the two languages most commonly used by Filipinos to express their sentiments. The results and methodologies from previous studies [15,22] were used to study this critical issue. The findings of this study will help the Philippine government make wise decisions in allocating funds and devising vaccination rollout plans. A comparison of tweet classification results in the Philippines using RM is listed in Table 2, including the authors, classifier algorithm, and the results obtained.

Table 2. Comparison of Philippine tweet classification results using RM and Naïve Bayes classifier algorithm.

Authors	Classifier	Results	Reference
Samonte et al.	Naïve Bayes	66.67% accuracy	[15]
Abisado et al.	Multinomial Naïve Bayes	72.17% accuracy	[22]
Proposed method	Naïve Bayes	81.77% accuracy	

The researchers made use of the study by Samonte et al. titled "Sentiment and Opinion Analysis on Twitter about Local Airlines", which compared Naïve Bayes, support vector machine, and random forest to develop a classifier model using RM to recognize the true polarity of tweets and concluded that the Naïve Bayes is the best among the three in terms of accuracy (66.67%) [15]. The same study was cited in the study by Abisado et al. titled "Philippine Twitter Sentiments during COVID-19 Pandemic Using Multinomial Naïve Bayes", which yielded an accuracy of 72.17%. Using RM and the Naïve Bayes classifier algorithm, the proposed method in this study obtained 81.77% accuracy, which is the highest in terms of accuracy.

3. Materials and Methods

To achieve the objectives of this study, the researchers started by collecting related tweets, followed by data annotation, data processing through NLP techniques, sentiment classification using Naïve Bayes classifier algorithm, and performance evaluation by applying the developed model in an unlabeled dataset. The approach of this study is displayed using a block diagram in Figure 1.

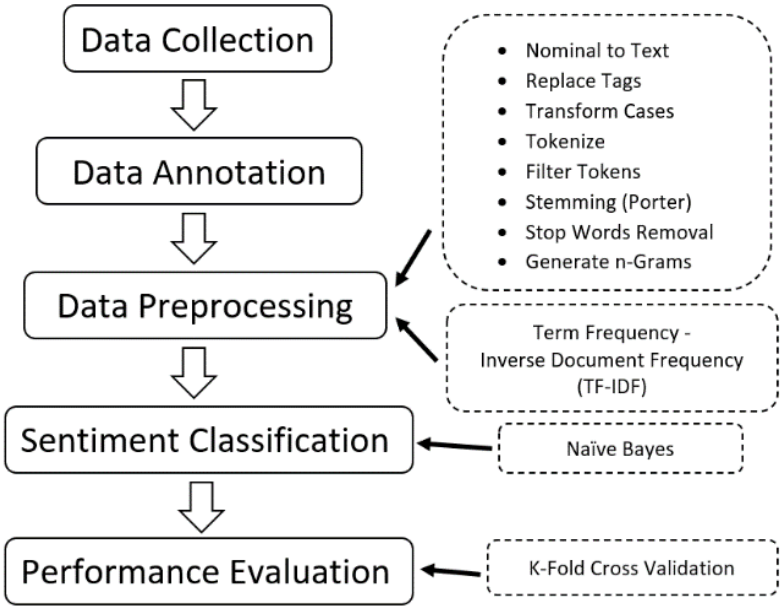


Figure 1. This figure presents the five phases of this study: data collection, data annotation, data preparation, sentiment classification, and performance evaluation.

Five phases were used in this study. In the data collection phase, tweets related to COVID-19 vaccines in the Philippines were gathered. After some data cleaning, such as removing retweets and duplicates, the data annotation phase was conducted, where the collected tweets were classified into positive, neutral, and negative polarities. The next and very important phase of the proposed method wherein the researchers spent a lot of time was the data preparation and preprocessing phase, which used several NLP techniques. This process helped to prepare the data for the next phase, which was to train a sentiment classification model using the Naïve Bayes classification algorithm. Lastly, the k-fold cross validation operator was used to evaluate the performance of the developed model. Detailed explanation of each phase of the methodology is given in Sections 3.1–3.5.

3.1. Data Collection

The RM Search Twitter operator was used to search for tweets particularly in the Philippines with the languages English and Tagalog. This process was done on a weekly basis from 1 to 31 March 2021, which was the first month of the vaccination program of the Philippine government. Hashtags such as #covidvaccineph, #covid19vaccineph, #resbakuna, #BIDABakunation, #BIDASolusyon, #WeHealAsOnePH, and #covaxph were used to search for relevant tweets. A total of 11,974 tweets were gathered. Unnecessary data such as duplicate tweets and retweets were removed using the Remove Duplicates and Filter Examples operators by excluding tweets with "RT" in the text. These processes were applied to minimize data redundancy and make the data set cleaner.

The researchers noticed that most of the gathered tweets were news items, where the sentiment of Filipino citizens towards COVID-19 vaccines was not emphasized. Therefore, tweets from news media companies were also removed. A total of 993 tweets were left and stored in an Excel file. Three types of operators were used in the data collection phase, namely the Search Twitter, Append, and Write Excel operators. The design of this process is displayed in Figure 2.

As shown in Figure 2, seven RM Search Twitter operators with the researcher's twitter account connection source were used to extract tweets containing the mentioned hashtags. The output of each operator was loaded as examples (exa) in the RM Append operator. The Append operator's job is to merge (mer) all the collected tweets and provide it as an input to the Write Excel operator to generate an Excel file where all the gathered tweets are stored for further processing and analysis.

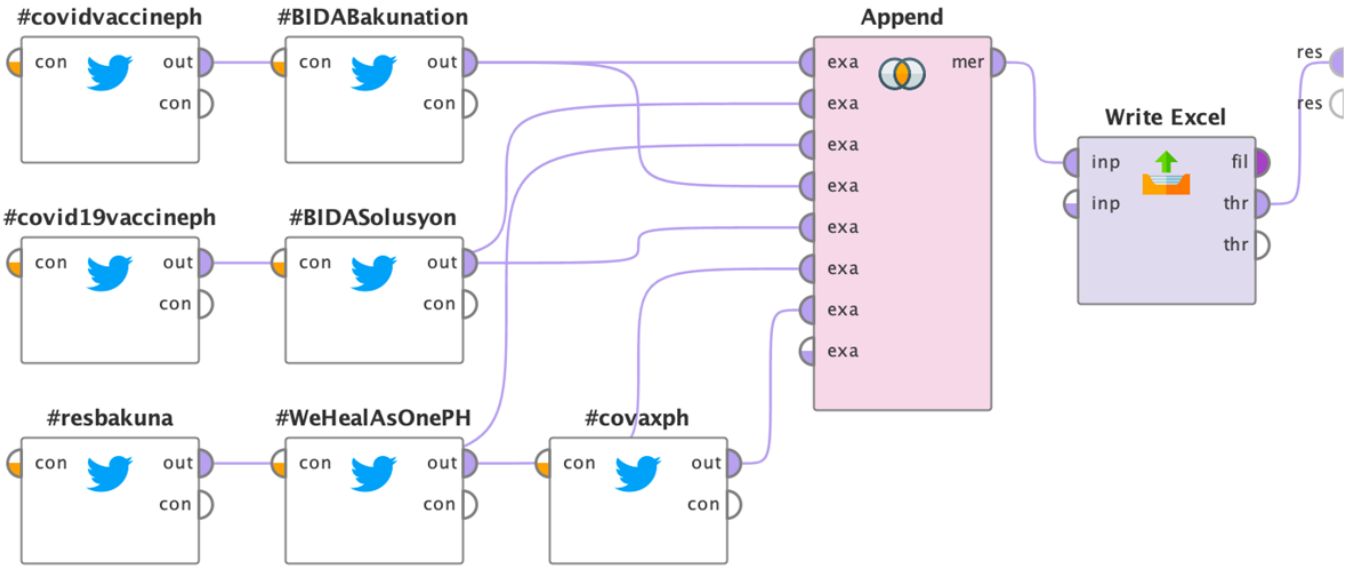


Figure 2. The RM Search Twitter operator with respective hashtags was used to collect tweets related to COVID-19, the Append operator was used to merge all the gathered tweets, and the Write Excel operator was used to store the data in an Excel file.

The cleaned tweets stored in the Excel file were imported in the RM local repository using the Import Data function. RM Retrieve operator was used to access the stored information in the dataset and load them into the data preprocessing operators. Initially, the attributes collected were date of creation, username, user ID, to user, language, source, text, geolocation latitude and longitude, retweet count, and tweet ID; some attributes unrelated to the study were removed. Two attributes were utilized in this study, namely date created and text. The date created attribute was used to process the data through time, while the text attribute contained the tweets to be analyzed in this study. An additional special attribute of polarity was created to categorize the tweets. An example set of cleaned tweets is shown in Table 3.

Table 3. These are some of the collected tweets imported as an example set in the RM application. Displayed attributes are date of creation, polarity, and the text of the tweets.

Date of Creation	Polarity	Text
8 March 2021	Positive	Have confidence! The doctor is (vacc)in(ated). UPLB alumna Dr. Sharon Madriñan-Garcia receives a first dose of Sinovac vaccine at the Ospital ng Palawan. #Bakunado #Resbakuna #KasanggaNgBida #ScienceWorks #VaccinesWork #VaccinationSavesLives
8 March 2021	Positive	1st dose done! I was expecting it to be painful & be getting that swell-feels after but there wasn't. Glad to be vaccinated after 1 year of covid19 exposure. I hope & pray this all ends so we can watch @BTS_twt again #RESBAKUNA #vaccinated #GetVaccinated #HealthIsWealth
8 March 2021	Positive	H O P E Got my first dose today. #RESBAKUNA #BIDABakunation #BIDASolusyon+ #VaccinesWork
8 March 2021	Positive	Got my first dose of Sinovac vaccine. #VaccineWorks #Resbakuna
8 March 2021	Positive	Get vaccinated! I got my Covid19 vaccine #Resbakuna #Sinovac #1stdose @ Jose B. Lingad Memorial General Hospital

Table 3. *Cont.*

Date of Creation	Polarity	Text
8 March 2021	Positive	#ResBAKUNA with SinoVac done.
8 March 2021	Positive	SINOVAC Vaccination todayyy #RESBAKUNA #SINOVAC
8 March 2021	Positive	1st dose done #RESBAKUNA #vaccinated
8 March 2021	Positive	Got my 1st shot today. #RESBAKUNA
8 March 2021	Neutral	today pala is the 1st day of vaccination sa bpmc #resbakuna
8 March 2021	Positive	* gets the @UniofOxford @AstraZeneca vaccine * Side effects include a temporary British accent lasting for a few hours. Me: Bloody hell that was painful! But cheers for the vaccine, mate! #VaccinesWork #RESBAKUNA

In Table 3, the date of creation of the tweet, polarity and text of the particular tweet were displayed, these are the raw data directly from the twitter search operator. Special symbols such as hashtags (#), mention or at tags (@), and asterisks (*) were still included.

3.2. Data Annotation

The tweets collected during the data collection phase were printed and annotated by hand with the aim of classifying the tweets into three polarities: positive, neutral, and negative. Positive polarity referred to tweets that were highly enthusiastic about the arrival of COVID-19 vaccines in the country. Tweets where the user showed willingness to be vaccinated and hoped that vaccines would work were marked as positive. Neutral polarity entailed tweets where the user either agreed or disagreed with the immunization program. Tweets where users expressed that they may or may not consider vaccination were marked as neutral. Negative polarity was given to tweets that totally disagreed with COVID-19 vaccines. Tweets that showed negative reactions, arguments, and refusal to the vaccines and included displeasure over adverse side effects after vaccination were marked as negative. Some manually annotated tweets belonging to the three classifications are shown in Table 4.

Table 4. Some examples of tweets with positive, neutral, and negative polarities that were annotated by hand and imported as a dataset in RM for data processing, model training, and testing.

Polarity	Tweet
Positive	I find it very comforting to see that we are not just counting COVID-19 cases but also the number of people getting vaccinated. Bright days are coming. #RESBAKUNA
	Felt an overwhelming sense of hope today. Vaccine saves lives. Praying for a brighter tomorrow. #ResBakuna #COVID-19Vaccine
	Anyone can be a HERO but getting vaccinated can make you SUPER! Finally received my first dose of the vaccine, today!
Neutral	I RESPECT the Philippine FDA recommendation that Sinovac is NOT for healthworkers. I also RESPECT the healthworkers who did NOT RESPECT the FDA recommendation by getting vaccinated with Sinovac #DuktorDapatAngHUWARAN #RESBAKUNA
	This right is never different from what the feminists are trying to advocate: Our body, our choice. #MyBodyMyChoice
	Hello guys sino na sa inyo ang nakapagpa bakuna? #RESBAKUNA

Table 4. *Cont.*

Polarity	Tweet
Negative	2nd day. Sama pakiramdam ko pra akong ttrangkasuhin, feeling sleepy and uhaw na uhaw #astrazenecavaccine #firstdose #RESBAKUNA
	Feverish. Headache. Hungry. It was so weird that this vaccine made me feel so hungry. Imagine feeling sick, but hungry. #resbakuna
	Karamihan sa frontline healthcare workers sa Palawan ang tumangging magpaturok sa COVID-19 vaccine na gawa ng Sinovac at piniling hintayin ang bakuna ng AstraZeneca.

The polarity column displays the classification of the tweet as positive, neutral, or negative, while the Tweet column displays some annotated tweets that belong to a particular category.

3.3. Data Preparation and Preprocessing

The learning operators available in RM are not only useful for data gathering but can also provide an easy framework for text mining and sentiment analysis. After the data annotation phase, the dataset was ready to be used to develop a classifier model. With the aim of achieving a classifier with high accuracy, the researchers used the available NLP techniques to clean the tweets to be used for training and testing. These techniques have also been utilized in other studies for sentiment analysis of English tweets [13,15,22]. Below are the step-by-step processes carried out in the data preparation and preprocessing phases.

3.3.1. Conversion of Nominal Values to Text

The Select Attributes operator was used to obtain the necessary attributes and remove other attributes [21]. In this study, a subset of attributes from the retrieved example set was selected, which included the text and the polarity of tweets. The Nominal to Text operator was used to convert the selected attributes to their corresponding text values. The nominal data type was used to name, label, and categorize the attributes without specific ordering, such as name of a person or thing, gender, and phone number [25]. The data type of the selected attributes was polynomial, which meant the data included a number of nominal values. Because the RM Process Document operator only accepts text inputs, this attribute needed to be converted to text data type. The output of this process was sent to the RM Process Document operator for NLP, which included case transformation, tokenization, stemming, stop words removal, and generating n-grams. The Select Attributes operator and the Nominal to Text operator are displayed in Figure 3.

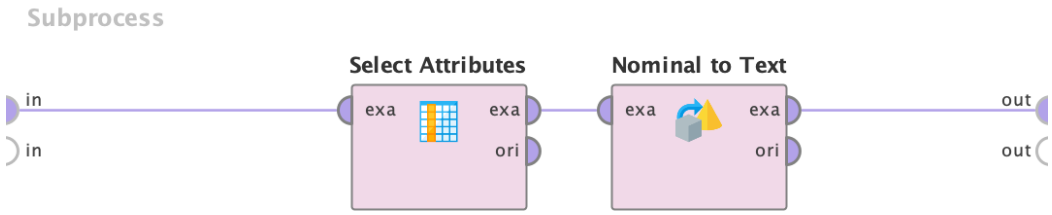
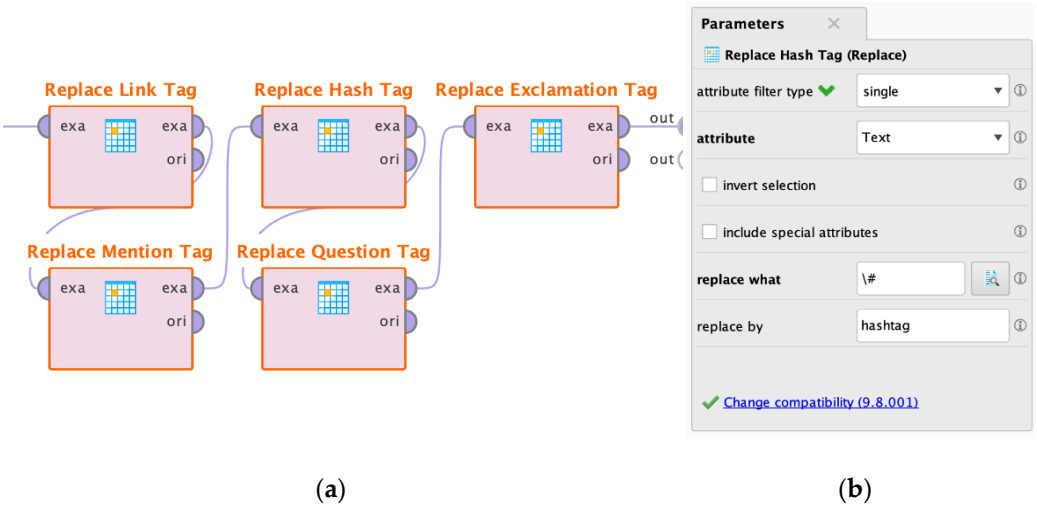


Figure 3. The input port (in) contained the entire example set to be processed in the Select Attributes and Nominal to Text operators. Converted examples were sent to the output port for preprocessing.

3.3.2. Replacing Tags

The RM Replace Tag operator was used to further clean the tweets. This operator searches for a tweet containing special characters, such as hashtags (#), question mark (?), exclamation point (!), link tags (http), and mention tags (@), and replaces them with the corresponding text values to facilitate the extraction of words [11].

When a special character mentioned above was found, it will be replaced with the corresponding word; for example, the symbol "#" will be replaced by the word "hashtag". The same process is applied for question mark, exclamation point, link, and mention tags. This process is necessary to determine the special symbols that would be ignored by the operators because it has no significant meaning necessary for NLP operators. The Replace Tag operator and the parameters used are displayed in Figure 4.



(**a**) (**b**)

Figure 4. (**a**) Replace Tag operator that replaces special symbols to their corresponding words, and (**b**) the parameter section of the operator to input "replace what" and "replace by" values.

Examples (exa) of texts from the tweets were loaded and processed by the Replace Tag operator until each tweet had undergone all the replace processes. The processed tweets were transferred using the output port (out) to other operators for further preprocessing. The "single" attribute filter type was used for the text attribute, which was the only attribute to be processed. The "replace what" parameter included the symbols to look for and the "replace by" parameter included the corresponding words.

3.3.3. Process Documents from Data

The RM Process Documents from Data operator generates word vectors from string attributes using term frequency–inverse document frequency (TF–IDF), which is a value that entails the importance of a word in a given document. TF pertains to how often the word occurs in a document with respect to the number of words used in the document [21]. RM uses the following formula to calculate the TF value.

$$TF\ (T, D) = T/D \quad (1)$$

where T is the number of times the term or the word appears, and D is the number of words in the document. TF considers all the words in the document as equally important, while IDF scales up the unique words present in the document and lowers the value of commonly used terms like the stop words "is", "of", and "the" [26].

$$IDF(T) = N/D \quad (2)$$

This formula was used to measure the importance of the term T, where N is the count of corpus or the overall number of words, and DF is the number of occurrences of the term T in the document. Zero cannot be a denominator, so adding 1 to it will avoid dividing by zero, which results in the following formula [26]:

$$IDF(T) = \log\ (N/(DF + 1)) \quad (3)$$

The final formula to measure how important a word is in a set of documents is as follows:

$$\text{TF-IDF (T, D)} = \text{TF (T, D)} \times \log (\text{N}/(\text{DF} + 1)) \quad (4)$$

The RM Process Documents from Data operator performs several processes to prepare the dataset to be used in developing a model for sentiment analysis. The following are the processes applied to the dataset.

1. Transform cases: This process transforms all the uppercase letters into lowercase and vice versa. The researchers chose to transform to lowercase in the parameter section.
2. Tokenization: This process splits the tweets into a sequence of tokens or terms. It also removes punctuation marks and white spaces present in the tweet [13,17].
3. Filter tokens by length: This process traverses throughout the tokenized terms and filters words shorter or longer than a specified number of characters. The researchers used a minimum of 4 and a maximum of 25 characters per word.
4. Stemming: This process uses the Porter stemming algorithm to replace suffixes present in words [17]. By doing so, the process can extract the root word in a term to obtain higher accuracy. For example, the words high, higher, and highest can all be stemmed to the word "high" [11].
5. Stop words removal: In this process, stop words were removed. The Filter Stop Words English operator was used to process the English tweets. Words such as the, a, an, with, of, etc. were removed. Tagalog stop words, such as ang, at, kay, na, o, din, ba, etc., were also removed using a text file containing Tagalog stop words as an input to the Filter Stop Words Dictionary operator. These two filter operators were used to cater for both English and Tagalog tweets present in the dataset.
6. Generate n-grams (Terms): The term n-gram means consecutive terms of a specified length n [15]. The researchers took into consideration a maximum of three consecutive words. The preprocessing techniques utilized in this study are shown in Figure 5.

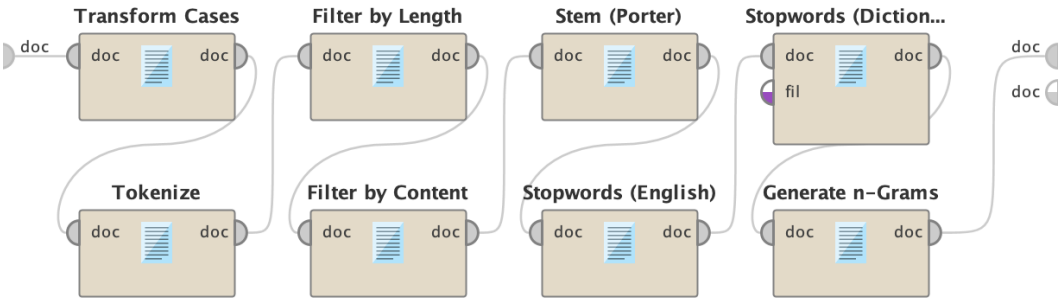


Figure 5. Subprocesses of the Process Document from Data operator. The inputs of the processes were the collected tweets that had undergone the data annotation phase, and the outputs were the word vectors.

Each tweet traversed through the eight operators in Figure 5 one step at a time. The Process Documents from Data operator created word vectors that were loaded as inputs for the training and testing dataset in building the sentiment classifier model.

3.4. Sentiment Classification

The Naïve Bayes classification algorithm was used to classify the tweets according to their polarity. As mentioned in related literature [22], Naïve Bayes works very well even in small datasets. Because the data collection was carried out during the first month of the implementation of the vaccination program, a total of 993 tweets were left after a thorough data cleaning process. Here, Naïve Bayes works best in classifying the true polarity of tweets. The Naïve Bayes classifier is based on Bayes' theorem, which is defined as follows:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)} \quad (5)$$

where, *P (A)* is the probability of hypothesis being true (regardless of data) or prior probability, *P (B)* is the probability of the data or evidence (regardless of probability) or marginal probability, *P(A | B)* is the probability of hypothesis A given data B or posterior probability, *P(B | A)* is the probability of data B given that hypothesis A is true or the likelihood probability. Because Naïve Bayes is independent of the occurrence of other features or in this case the data, we assume *B* to be *B1*, . . . , *Bn*, which gives the result as follows:

$$P(A|B_1 \dots B_n) = \frac{P(B_1|A)P(B_2|A) \dots P(B_n|A)P(A)}{P(B_1)\ P(B_2) \dots\ P(B_n)} \tag{6}$$

which can be expressed as

$$P(A|B_1 \dots B_n) = \frac{P(A)\prod_{i=1}^{n} P(B_i|A)}{P(B_1)\ P(B_2) \dots\ P(B_n)} \tag{7}$$

Because the denominator is constant all throughout a given input, the term can be removed:

$$P(A|B_1 \dots B_n) \propto P(A)\prod_{i=1}^{n} P(B_i|A) \tag{8}$$

For the classifier model, it is important to find the probability of the data given for all possible values of *A* and the output with maximum probability, which can be expressed mathematically as follows:

$$A = argmax_A P(A)\prod_{i=1}^{n} P(B_i|A) \tag{9}$$

Finally, the researchers were left with the task of calculating *P(A)*, which is called the class probability, and *P(B_i | A)*, which is the conditional probability [23]. The classifier calculates the probability of an event in the following steps [24]:

Step 1: Calculate the prior probability for given class labels.

Step 2: Find likelihood probability with each attribute for each class.

Step 3: Put these values in Bayes formula and calculate posterior probability.

Step 4: See which class has a higher probability given the input belongs to the higher probability class.

There is an available operator named "Naïve Bayes" in RM under the modeling and predictive directories, which can be used to perform training to develop a model to classify tweets according to their polarity. This operator can be seen in Figure 6.

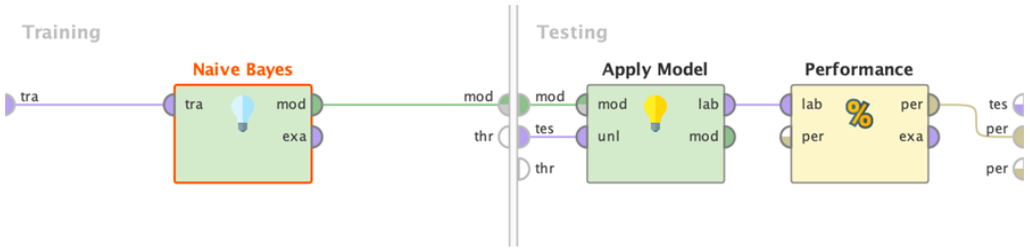


Figure 6. Subprocesses of the Cross Validation operator displaying the training section using Naïve Bayes classifier algorithm and testing sections displaying the model's performance.

3.5. Performance Evaluation

To evaluate the statistical performance of the model, the researchers applied the RM K-fold Cross Validation operator that has a single parameter K, which is the number of groups where the data can be split into training and testing examples. The researchers used 10 folds, which technically means that the evaluation tool used was a 10-fold cross validation. The subprocesses of the Cross Validation operator is shown in Figure 6.

As shown in Figure 6, the researchers used automated mode to select preprocessed tweets to allot data for the training and testing datasets. The left section displays the training process using the Naïve Bayes classification algorithm to build a classifier model. On the right is the testing process, which makes use of the other set of tweets not used in the training phase to fairly evaluate the performance of the developed model. The Apply Model operator was used for the classification process in the test dataset or the unlabeled (unl) data. The training continued until the operator finished labelling all the data. The labelled data were loaded to the Performance operator to evaluate the performance of the model. The output of the Performance operator was the performance (per) port, which carried information regarding the result of the developed model.

4. Results and Discussion

A total of 11,974 tweets were collected using the RM Search Twitter operator, and a total of 993 tweets were left after data preparation and preprocessing. Among the 993 tweets, 828 or 83.38% of the collected tweets pertained to a positive attitude. In most of the tweets, the users mentioned that they chose to be vaccinated to encourage others to do the same. Phrases like "trust science", "vaccine works", and "a dose of hope" were observed several times.

Few tweets having negative polarity were also analyzed. Some tweets included adverse side effects, such as fever, chills, and dizziness, or noted that the area where the vaccine was injected was a little heavy and sore. However, some tweets mentioned that it was fine because the benefit of having a vaccine outweighs the risk of having the virus. Therefore, only 82 or 8.26% of the tweets had negative polarity. Likewise, 83 or 8.36% of the tweets had neutral polarity. A graphical visualization of the processed dataset according to polarity is shown in Figure 7.

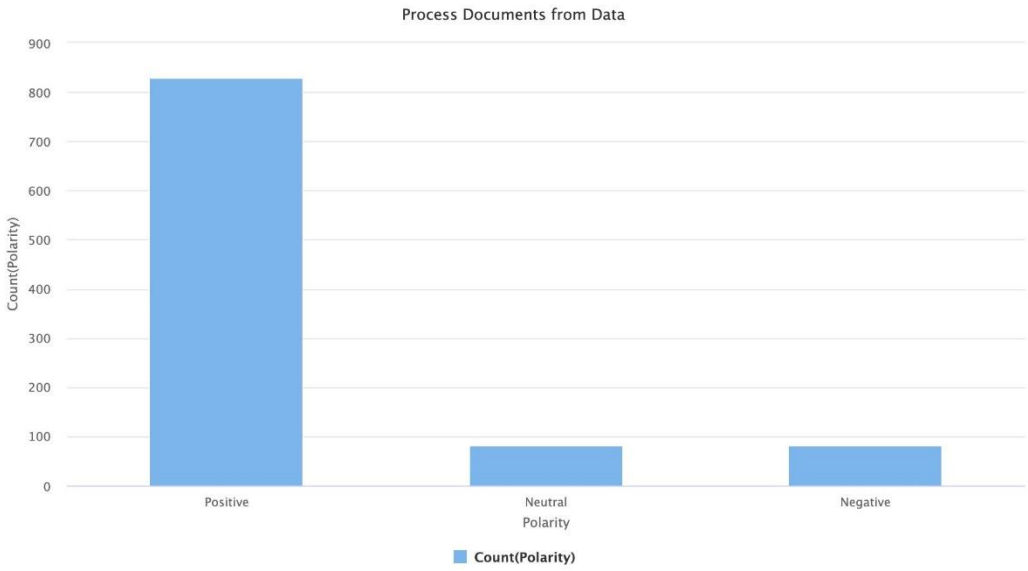


Figure 7. Processed dataset according to their classification as positive, neutral, or negative.

The annotated data were utilized to train a classifier model using the Naïve Bayes classification algorithm. After training and testing, the results showed that the model attained 81.77% accuracy. The confusion matrix, which included the true positive, neutral, and negatives and the number of predicted items, is shown in Table 5. Precision and recall percentages were also calculated.

Table 5. Confusion matrix, which is the result of the Performance operator after applying the developed model to training or unlabeled dataset.

Label.	Predicted Positive	Predicted Neutral	Predicted Negative	Class Precision
True Positive	**745**	57	26	89.98%
True Neutral	47	**29**	6	35.37%
True Negative	38	7	**38**	45.78%
Class Recall	89.76%	31.18%	54.29%	

In Table 5, the columns represent the developed model's predicted positive, predicted neutral, and predicted negative polarities, while the last column displays the percentage of the class precision. On the other hand, the rows represent the true or the actual positive, neutral, and negative tweets from the dataset, while the last row pertains to the class recall percentage. Based on the results, the developed model successfully predicted 745 positive tweets among the 828 labeled true positive tweets, which equates to class precision of 89.98% and class recall of 89.76% for the positive polarity. The numbers in bold format indicates the number of correctly predicted polarities. Thus, 745, 29 and 38 were the correctly predicted tweets for positive, neutral and negative tweets respectively.

The researchers wanted to analyze the preliminary sentiments towards COVID-19 vaccines in the Philippines in the first month of the implementation of the vaccination program. Figure 8 presents a graphical visualization of the tweets per day from 1 to 31 March 2021, grouped according to their polarity through time. There were only a few tweets on the first day of implementation of the COVID-19 vaccination program, but this eventually increased until reaching the highest number of tweets on 8 March 2021, exactly one week into the implementation. Then, the tweets decreased on the second week of March until the third week, when the number of tweets began to rise again, primarily because of the arrival of the fresh batch of Sinovac doses donated by China.

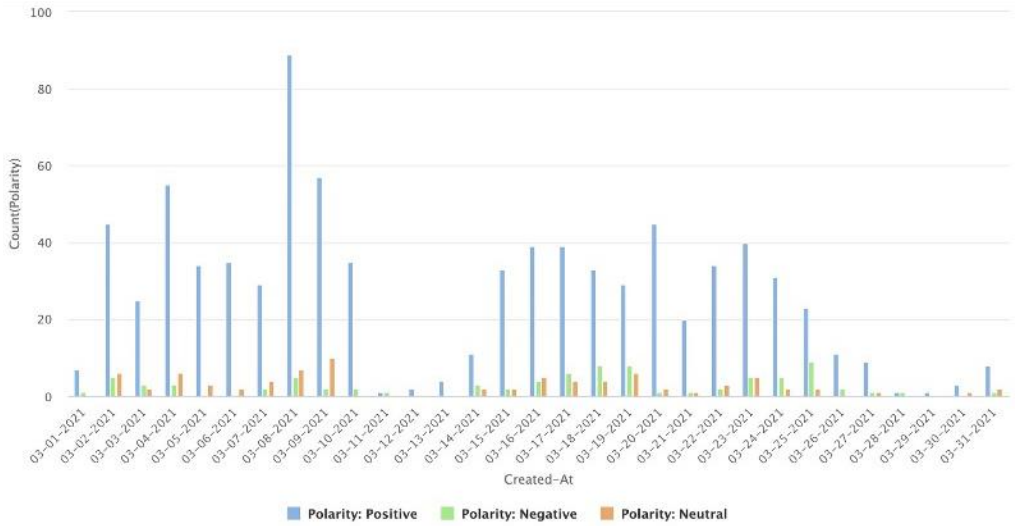


Figure 8. Processed dataset through time aggregated by the polarity of tweets. The chart shows tweets per day and their polarity classified as positive, negative, and neutral.

In the bar graph, the horizontal axis represents the date in which the tweet was posted, and the vertical axis represents the number of tweets. Tweets per day were aggregated according to their polarity. It is evident from the graph that the number of positive tweets, indicated by the blue bars, were the highest for each day. Neutral and negative tweets, indicated by orange and green bars, respectively, were very minimal. The words and the number of times they appeared per classification are displayed in Table 6, and

the corresponding word cloud for positive, neutral, and negative classification is shown in Figure 9.

Table 6. Word count per classification.

Positive		Neutral		Negative	
Word	**Count**	**Word**	**Count**	**Word**	**Count**
resbakuna	655	resbakuna	58	vaccine	74
vaccine	645	vaccine	56	covid	62
covid	424	covid	41	resbakuna	34
bidabakunation	223	sinovac	14	astrazeneca	13
bidasolusyon	217	bidabakunation	13	bakuna	12
explain	162	bidasolusyon	13	hindi	11
dose	145	explain	11	vaccinated	11
astrazeneca	124	health	9	covidvaccineph	10
sinovac	99	astrazeneca	7	sinovac	10
bakuna	96	bakuna	7	health	9
shot	51	city	6	kaya	9
magpabakuna	50	respect	6	explain	8
health	49	dose	5	workers	8
medical	48	getting	4	bidasolusyon	7
thank	46	healthworkers	4	people	7

(**a**) (**b**) (**c**)

Figure 9. Word cloud: (**a**) positive polarity, (**b**) neutral polarity, and (**c**) negative polarity. The word cloud displays the words used in the dataset; the more frequently the word was used, the bigger it is displayed in the word cloud.

5. Conclusions

This study aimed to analyze sentiments towards COVID-19 vaccines in the Philippines according to positive, neutral, and negative polarities. Based on the results, it can be concluded that the majority or 83% of the tweets in the Philippines were positive and enthusiastic about the idea of vaccination, while 9% had neutral and 8% had negative sentiments. The data were preprocessed using several NLP techniques, and a classifier model was successfully developed using the Naïve Bayes classification algorithm with 81.77% accuracy through RM operators. Because the Naïve Bayes works very well even in a small dataset, it was used for this study, which was composed of tweets for the first month of the vaccination program in the Philippines.

Sentiment analysis towards COVID-19 vaccines can help the Philippine government make wise decisions regarding allocation of funds and vaccination rollout plans. The developed model using the Naïve Bayes classification algorithm can help classify tweets according to their polarity, especially for English and Tagalog languages.

The limitation of this study is that due to the immediate implementation of COVID-19 vaccination in the country, the researchers made use of the tweets on the first month, which was 1–31 March 2021. However, as time passes, additional tweets can be easily gathered and automatically labelled according to their polarity to add more training and testing data. The free version of RM has some limitations in terms of the number of data or rows that can be imported, and the subscription rate is not cheap. The process of sharing the results, analysis, data, and graphs generated from RM needs to be improved. Moreover, the researchers used the available operators in the RM that do generate codes based on the developed model [27]. The models under the Extract Sentiment operator in RM do not include the Filipino language, so the researchers had to annotate the gathered tweets by hand.

The researchers recommend continuous analysis of the sentiments of people towards COVID-19 vaccines. Studies that aim to determine if there is a significant correlation between the number of vaccinated people, active cases, and number of deaths compared to the time when these vaccines were not yet available should also be considered. Future research can also classify tweets into different emotions, such as inspired, happy, annoyed, sad, angry, afraid, etc. to fully understand and reveal the sentiment of the tweets.

The journey of winning the battle against COVID-19 still has a long way to go. However, having effective vaccines and citizens willing to be vaccinated is a great move towards achieving this goal.

Author Contributions: Conceptualization, C.V. and J.J.M.; methodology, C.V. and J.J.M.; validation, J.J.M. and X.A.I.; formal analysis, C.V. and J.J.M.; writing—original draft preparation, C.V.; writing—review and editing, J.J.M., X.A.I., and J.-H.J.; visualization, C.V.; supervision, J.-H.J. and J.-G.H.; project administration, J.-H.J. and J.-G.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Bulacan State University as an International Faculty Scholar and I-Shou University together with Taiwan Ministry of Education (MOE) as MOE New Southbound Elite Scholar.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Acknowledgments: The authors wish to thank Bulacan State University, Philippines, and I-Shou University, Taiwan. This work was supported in part by a grant from Bulacan State University as an International Faculty Scholar and Taiwan's Ministry of Education (MOE). The researchers would also like to thank the ALMIGHTY GOD for His guidance from the start until the completion of this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jan, A.; Mata, M.N.; Albinsson, P.A.; Martins, J.M.; Hassan, R.B.; Mata, P.N. Alignment of Islamic Banking Sustainability Indicators with Sustainable Development Goals: Policy Recommendations for Addresing the COVID-19 Pandemic. *Sustainability* **2021**, *13*, 2607. [CrossRef]
2. Galea, S.; Merchant, R.M.; Lurie, N. The Mental Health Consequences of COVID-19 and Physical Distancing the Need for Prevention and Early Intervention. *JAMA Intern. Med.* **2020**, *180*, 817–818. [CrossRef] [PubMed]
3. Tirachini, A.; Cats, O. COVID-19 and Public Transportation: Current Assessment, Prospects, and Research Needs. *J. Public Transp.* **2020**, *22*. [CrossRef]
4. Sharma, S.; Sharma, M.; Singh, G. A chaotic and stressed environment for 2019-nCoV suspected, infected and other people in India: Fear of mass destruction and causality. *Asian J. Psychiatry* **2020**, *51*, 102049. [CrossRef] [PubMed]
5. Madarang, C.R.S. COVID-19 Vaccines Reach All Southeast Asian Nations except Philippines. *Interaksyon*, 26 February 2021. Available online: https://interaksyon.philstar.com/politics-issues/2021/02/26/186349/covid-19-vaccines-reach-all-southeast-asian-nations-except-philippines/ (accessed on 7 April 2021).
6. Tomacruz, S. Philippines Receives First COVID-19 Vaccine Delivery from Sinovac. *Rappler*, 28 February 2021. Available online: https://www.rappler.com/nation/philippines-receives-first-delivery-covid-19-vaccine-sinovac-february-28-2021 (accessed on 8 April 2021).

7. Reuters. Philippines to Continue AstraZeneca Vaccinations Amid Suspensions in Europe. *Reuters*, 12 March 2021. Available online: https://www.reuters.com/article/us-health-coronavirus-philippines-vaccin-idUSKBN2B403X (accessed on 8 April 2021).
8. De Leon, D. Philippines Receives 400,000 More Sinovac Doses from China. *Rappler*, 24 March 2021. Available online: https://www.rappler.com/nation/philippines-receives-more-china-sinovac-vaccines-march-24-2021 (accessed on 8 April 2021).
9. Venzon, C. Philippines Starts COVID Vaccinations, Courtesy of China. *Nikkei Asia*, 1 March 2021. Available online: https://asia.nikkei.com/Spotlight/Coronavirus/COVID-vaccines/Philippines-starts-COVID-vaccinations-courtesy-of-China (accessed on 8 April 2021).
10. Department of Health (Philippines). The Philippine National Deployment and Vaccination Plan for COVID-19 Vaccines. January 2021. Available online: https://doh.gov.ph/sites/default/files/basic-page/The%20Philippine%20National%20COVID-19%20Vaccination%20Deployment%20Plan.pdf (accessed on 8 April 2021).
11. Sharma, S. GitHub. 8 May 2018. Available online: https://github.com/stuti-sharma/Sentiment-Analysis-Twitter-RapidMiner (accessed on 5 April 2021).
12. Ndasauka, Y.; Hou, J.; Wang, Y.; Yang, L.; Yang, Z.; Ye, Z.; Hao, Y.; Fallgatter, A.J.; Kong, Y.; Zhang, X. Excessive use of Twitter among college students in the UK: Validation of the Microblog Excessive Use Scale and relationship to social interaction and loneliness. *Comput. Hum. Behav.* **2016**, *55*, 963–971. [CrossRef]
13. Cureg, M.Q.; De La Cruz, J.A.D.; Solomon, J.C.A.; Saharkhiz, A.T.; Balan, A.K.D.; Samonte, M.J.C. Sentiment Analysis on Tweets with Punctuations, Emoticons, and Negations. In Proceedings of the 2019 2nd International Conference on Information Science and Systems, Tokyo, Japan, 16–19 March 2019.
14. Guo, X.; Li, J. A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019.
15. Samonte, M.J.C.; Garcia, J.M.R.; Lucero, V.J.L.; Santos, S.C.B. Sentiment and opinion analysis on Twitter about local airlines. In Proceedings of the 3rd International Conference on Communication and Information Processing, Tokyo, Japan, 24–26 November 2017.
16. Liu, B. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*; 2010; Available online: https://www.cs.uic.edu/~{}liub/FBS/NLP-handbook-Liub-final.pdf (accessed on 9 May 2021).
17. Basiri, M.E.; Nemati, S.; Abdar, M.; Cambria, E.; Acharrya, U.R. ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Gener. Comput. Syst.* **2021**, *115*, 279–294. [CrossRef]
18. Wei, L.; Wei, S.; Ji, S.; Cambria, E. BiERU: Bidirectional Emotional Recurrent Unit for Conversational Sentiment Analysis. *arXiv* **2020**, arXiv:2006.00492.
19. Ali, F.; Ali, A.; Imran, M.; Naqvi, R.A.; Siddiqi, M.H.; Kwak, K.-S. Traffic accident detection and condition analysis based on social networking data. *Accid. Anal. Prev.* **2021**, *151*, 105973. [CrossRef] [PubMed]
20. Ali, F.; Sappagh, S.; Islam, S.R.; Ali, A.; Attique, M.; Imran, M.; Kwak, K.-S. An Intelligent Healthcare Monitoring Framework using Wearable Sensors and Social Neworking Data. *Future Gener. Comput. Syst.* **2020**, *114*, 23–43. [CrossRef]
21. Rapid Miner Documentation. Rapid Miner. 2021. Available online: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/bayesian/naive_bayes.html (accessed on 5 April 2021).
22. Delizo, J.P.D.; Abisado, M.B.; De Los Trinos, M.I.P. Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 408–412.
23. Khurana, S. Naive Bayes Classifiers Geeks for Geeks. 15 May 2020. Available online: https://www.geeksforgeeks.org/naive-bayes-classifiers/ (accessed on 3 May 2021).
24. Navlani, A. Naive Bayes Classification using Scikit-learn. 5 September 2020. Available online: https://avinashnavlani.medium.com/naive-bayes-classification-using-scikit-learn-60bc5176f868 (accessed on 4 May 2021).
25. Form Plus Blog. 10 December 2020. Available online: https://www.formpl.us/blog/nominal-ordinal-interval-ratio-variable-example#:~{}:text=A%20nominal%20variable%20is%20a,intrinsic%20ordering%20of%20these%20categories.&text=Some%20examples%20of%20nominal%20variables,%2C%20Name%2C%20phone%2C%20etc (accessed on 27 April 2021).
26. Hamdaoui, Y. TF(Term Frequency)-IDF(Inverse Document Frequency) from Scratch in Python. 10 December 2019. Available online: https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558 (accessed on 5 May 2021).
27. Pros and Cons of Rapid Miner Studio 2021. 2021. Available online: https://www.trustradius.com/products/rapidminer-studio/reviews?qs=pros-and-cons (accessed on 6 May 2021).