SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser
p-ISSN 2964-1160 | e-ISSN 2963-6361

JOISER
Journal of Information System Exploration and Research
Vol 01, No 01, January 2023

Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest

Jumanto[1*], Much Aziz Muslim[2], Yosza Dasril[3], Tanzilal Mustaqim[4]
[1,4]*Department of Computer Science, Universitas Negeri Semarang, Indonesia*
[2]*Postgraduate Student, Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia*
[3]*Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia*

DOI: https://doi.org/10.52465/joiser.v1i1.104
Received 07 December 2022; Accepted 09 December 2022; Available online 09 December 2022

Article Info

Keywords:
Sentiment analysis;
Random forest;
Vader;
Twitter

Abstract

This study conducted a sentiment analysis of the impact of the Covid-19 pandemic in the economic sector on people's lives through social media Twitter. The analysis was carried out on 23,777 tweet data collected from 13 states in Malaysia from 1 December 2019 to 17 June 2020. The research process went through 3 stages, namely pre-processing, labeling, and modeling. The pre-processing stage is collecting and cleaning data. Labeling in this study uses Vader sentiment polarity detection to provide an assessment of the sentiment of tweet data which is used as training data. The modeling stage means to test the sentiment data using the random forest algorithm plus the extraction count vectorizer and TF-IDF features as well as the N-gram selection feature. The test results show that the polarity of public sentiment in Malaysia is predominantly positive, which is 11,323 positive, 4105 neutral, and 8349 negative based on Vader labeling. The accuracy rate from the random forest modeling results was obtained 93.5 percent with TF-IDF and 1 gram.

1. Introduction

The Covid-19 pandemic has reached all regions of the world. Including Malaysia. When the covid-19 outbreak occurs, people are required to comply with government regulations, one of which is not to do activities in crowds and public places and are required to always be at home and always maintain health protocols. Community activities that are usually found in the public environment suddenly stop which of course affects many areas of personal life and the lives of many people. One of the fields in the field of economics. In the field of economy, in ordinary practice, it involves the interaction of many people, such as direct buying and selling and workers in production.

[*] ***Corresponding Author:***

Jumanto,
Department of Computer Scince,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: jumanto@mail.unnes.ac.id

Various kinds of public responses emerged to the covid-19 pandemic events. One of the factors affected is in the economic field. Since the lockdown was carried out, many economic conditions in the community have been directly affected. The effect in question is a negative effect such as the source of income not running as usual or even completely stopped, which causes the need to manage expenses appropriately. Public unrest can be captured and read by many people through many media platforms, especially social media [1]. The impact of the Covid-19 pandemic has also affected people in Malaysia [2]. The Malaysian government imposed a movement control order or lockdown on March 18, 2020, which requires the community not to carry out activities in public places and stay at home. This was done to prevent the spread of the Covid-19 outbreak from spreading. Many shops, shopping places, and places that support the Malaysian economy are closed, affecting the general economic condition [3].

Public sentiments are written clearly from the choice of words used and hashtags in tweets that are widely disseminated [4]. Communication carried out by the community on social media shows the emotions that are being felt. These emotions are expressed in writing and then disseminated to the public. For example, community unrest when undergoing self-quarantine at home is conveyed through writing on social media. Writing on social media is then responded to by other people who have the same emotional condition. This causes communication between levels of society with various points of view. The collection of public communication through social media shows hidden insights, especially those related to sentiment analysis [5]. Sentiment analysis is often used by previous researchers to clearly determine the public response to an event [6]. Sentiment analysis is the process of extracting the polarity of public sentiment through various media [7]. One of the media used in sentiment analysis is text media. Social media is a medium that is often used in sentiment analysis. People gather and communicate with each other as it is done in the real world. Data from social media can be used to analyze an event that is happening in the community.

The social media platform used in this research is Twitter. Twitter provides an API (Application Programming Interface) for retrieving tweets based on predefined criteria. The Twitter API makes it easy for researchers to collect data that fits the research theme. There have been many studies that use social media twitter as a medium for collecting research data. Al-Khalifa et al in 2020 detected hate speech on Twitter social media, whose language base is Arabic [8]. Research on stock price prediction conducted by Das et al in 2018 used tweet polarity data [9]. Sentiment analysis was carried out in the political field by Kušen and Strembeck in 2016 regarding the public response to the presidential election in Austria using the network science method [10]. There is also an analysis of sentiment regarding speeches delivered by Donald Trump's and Hillary Clinton's during the US presidential election process in 2016 using machine-based methods by Liu and Lei [11]. In 2018 Kumar & Harish conducted a research on sarcasm classification using a content-based feature selection method on the Amazon product review dataset [12].

In this study, using filters was limited to those related to economic factors during the Covid-19 pandemic outbreak. Economic factors were chosen because based on the discovery of public information via social media twitter during the research, many were dissatisfied with government performance, and tended to experience deficits from financial conditions due to reduced income and needs that had to be met. Sentiment analysis in this research went through several stages, namely pre-processing, labeling and modeling [13]. The pre-processing stage functions to collect tweet data from the Twitter API in accordance with the research criteria. The data that has been collected is then cleaned of data that does not have a significant effect on the results of the study, such as inappropriate punctuation and data preparation before entering the labeling and modeling stages. The labeling stage is to provide a sentiment assessment on each tweet data based on a dictionary available in the Vader lexicon polarity detection library [14]. Sentiment assessment is divided into three types, namely negative, neutral and positive. The process of giving sentiment labeling works by detecting the polarity of a sentence or words and showing the level of status to be positive, negative or neutral depending on the measurement results [15]. Previously, several researchers have used vader to conduct sentiment analysis, namely Al-Natour and Turetken, which analyzed star ratings on consumer review datasets in 2020 [16]. Alaei et al in 2019 conducted an analysis on travel topics that collected data from various internet sources and labeled sentiments based on the vader library [17]. Vader is used in this study as a library to support sentiment labeling as training data before entering the modeling process.

Data that has been labeled are processed and tested using machine learning algorithms. The machine learning algorithm used in this study is random forest. Random forest has many advantage criteria, it can process a lot of data that has incomplete attributes and can work when processing large amounts of data [18]. The research includes the sentiment analysis of amazon product reviews by Al-Amrani et al in 2018 combined with a machine learning algorithm support vector machine [19]. Sentiment analysis in the social sphere of society such as the anti-LGBT case in Indonesia was researched by Fitri et al in 2019, which one of the methods uses the machine learning random forest algorithm [20]. Parmar et al in 2014 conducted a sentiment analysis study using a random forest that had hyperparameters set to get better results than a standard random forest [21]. The results of this study are to conduct a new sentiment analysis method from a combination of vader lexicon polarity detection and random forest machine learning algorithms that can be used to analyze public sentiment towards the Covid-19 pandemic, especially in the broad economy. The rest of this paper is organized as follows. Section 2 describes the method of collecting and processing sentiment raw data into analysis data from Twitter. The results of the analysis data were processed using vader and random forest. Section 3 describes in detail the discussion of the research which begins with a general tweet analysis followed by a sentiment analysis of each province. Finally, Section 4 presents an overall conclusion from the research that has been done.

2. Method

This research is divided into three stages, namely pre-processing, labeling, and modeling.

2.2. Research Method

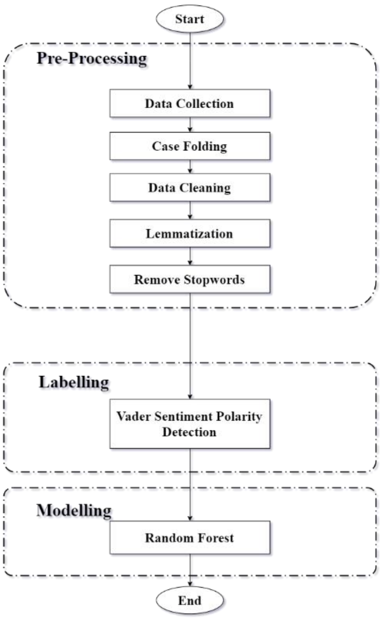

Figure 1. Flowchart research method

2.2.1. Pre-processing

The pre-processing stage is the process of collecting tweet data, the process of cleaning text data from some useless data and preparing the data before entering the main analysis process [22]. this needs to be done to prepare and clean the text from several attributes that make the main analysis process disrupted. Examples of things that need to be cleaned are such as irrelevant stopwords, uniform text forms, and abnormal text writing. The steps that can be taken during the pre-processing stage are case folding, data cleaning, lemmatization, and stop words. The data pre-processing process in this study was carried out with the NLTK library in the Python programming language.

2.2.2. Data Collection

The process of collecting data on Twitter social media uses the Python programming language with the GetOldTweet library. The GetOldTweet library provides the ability to retrieve tweet data in real-time on the Twitter platform based on predefined criteria [23]. The GetOldTweet library is used in this study because it has advantages, namely the time limit allowed for more than 7 days and even up to an annual period. In addition to the GetOldTweet library, you can also use the tweepy library, but data can only be retrieved within the last 7 days. The criteria that were used in the process of searching and collecting tweets were location factors that included all provincial capitals in Malaysia and predetermined search keywords related to the economy and covid-19. Search locations for the state capitals of Malaysia include the areas "Johar Bahru", "Alor Setar", "Kota Bharu", "KotaMelaka", "Seremban", "Kuantan", "Georgetown", "Ipoh", "Kangar", "Kota Kinabalu", "Kuching", "Shah Alam" and "Kuala Terengganu". The process of entering regional criteria in the GetOldTweet library based on coordinates: "1.5450255, 103.6395867", "6.1294823, 100.2834061", "6.1248062, 102.2456277", "2.2375488, 102.1814631", "2.712322, 101.9019532", "2.7126609, 101.901953 "," 5.4059643, 100.2743269 "," 4.6101207, 101.0215671 "," 6.4910077, 100.1823724 "," 5.9961432, 116.0254819 "," 1.6185192, 110.1859814 "," 3.0909411, 101.3764259 "and" 5.4845334, 102.764259 "and" 5.4845334, 102.764477997.

The keyword criteria included in this study are used in relation to the economy and covid-19. The list of keywords referred to is "money", "profesion", "fired", "muflis", "covid19", "covid-19", "pandemic", "profit", "loss", "entrepreneur", "buy "," commerce "," ringgit "," business "and" workers ". The criteria for the radius of the search area entered are as far as 50 miles from each of the coordinate centers of the provincial capital area. The results of data collection were obtained as many as 23777 tweets from all state capitals in Malaysia. Detailed data from all the capitals are Johar Bahru with 8640, Alor Setar with 1334, Kota Bharu with 732, Malacca City in 1931, Seremban with 2712, Kuantan with 2762, George Town with 970, Ipoh with 1953, Kangar with 1036, Kota Kinabalu. 639, Shah Alam 2793, and Kuala Terengganu 2797. The collected tweet data consists of 8 attributes, namely id_user, username, tweet text, date of the tweet, number of retweets, number of favorites, location of geo location, and number of hashtags used in tweets. Geolocation is used to filter the retrieval of tweets according to the criteria required in the study. Then the number of favorites and retweets is taken with the most dominant having the most number showing popularity to represent public opinion in tweets in the related search area.

2.2.3. Case Folding

The various forms of writing text made the analysis process difficult because the computer could not distinguish between upper and lowercase letters [24]. Writing uniform text serves to make the computational process more simplified and to improve the quality of the main analysis results.

2.2.4. Data Cleaning

Writing text from social media often contains random characters that affect the results of the main analysis [25]. examples are the character "@" and an incomplete emoji. Characters that affect the main analysis result can be cleaned using regular expressions or replace with influential characters in the analysis. Examples of replacement are from the character "@" can be replaced with "on". The process of cleaning text becomes one of the essences of pre-processing data.

2.2.5. Lemmatization

Lemmatization means making the text normal. Returning from outside the basic form to the basic form [26]. An example is "eating" changed to the basic form of "eating". This needs to be done to make the computation process simpler and then improve the quality of the main analysis.

2.2.5. Remove Stopwords

Stopword is the use of dominant words that are often used in speaking and writing [27]. Examples of using stopwords are "and","or","them" and "me". The use of a large-scale stopword makes the noise data uncontrollable which will ultimately affect the results of the main analysis.

2.2.6. Labelling

The labeling stage is the stage of giving a sentiment assessment label on tweet data [28]. The labeling process uses Vader sentiment polarity detection. Vader sentiment polarity detection works by matching the words found in the analysis process which are then matched with the polarity value in the predefined Vader dictionary. The labeling process is useful for knowing the sentiment results and is useful as a support for training data before proceeding to the modeling process using the random forest algorithm.

2.2.7. Vader sentiment analysis

Vader is an acronym for Valence Aware Dictionary for Social Reasoning which is used as a model for sentiment analysis and can determine the diversity of data through the intensity of emotional strength available according to the available Lexicon data dictionary [14]. Vader was introduced in 2014 by C.J Hutto and Eric Gilbert whose formation method is based on a human-centric approach, combining qualitative analysis and empirical validation using wisdom and human judgment [15]. Vader can provide a different polarity between "I like you" and "I don't like you". Polarity assessment combines lexical dictionary features with a sentiment score of 5 additional criteria namely exclamation mark, uppercase, word order level, polarity shift due to the word "but" and uses the tri-gram feature to check for the presence of negation [17]. The lexical approach aims to map words into sentiments by building a lexicon or 'dictionary of sentiments.

The Lexicon Dictionary can be used to rate the sentiment of phrases and sentences, without looking at the others. Sentiment can be categorized - such as {negative, neutral, positive} - or it can be numerical - such as a range of intensity or score. The lexical approach looks at the sentiment category or score of each word in a sentence and decides on the category or sentiment score of the whole sentence [29]. The strength of the lexical approach lies in the fact that it is not necessary to train the model using labeled data. Vader is an example of the lexical method. The advantage of using Vader polarity detection is that there is a dictionary that contains the value of each word. The process of determining the polarity of a sentence is obtained from the unifying attribute "compound" of each available word [30]. The criteria for grouping are positive, neutral, and negative, that is, if the compound result is more than 0.05, then the positive category is represented by the number 1 then if the compound result lies between -0.05 and 0.05 it is included in the neutral category represented by the number 0 and Finally, if the compound result is below -0.05 then it is a negative category which is represented by the number -1.

2.2.8. Modelling

The modeling stage is the process of analyzing the sentiment test analysis from the labeling stage using the random forest machine learning algorithm. The extraction feature complements the random forest to help the calculation process. The extraction features used in this study are the count vectorizer and TF-IDF. The selection of features in random forest uses N-gram. The final result of the modeling stage is the level of accuracy obtained from the random forest calculation process.

2.2.9. Random Forest

A random forest starts with a basic data mining technique, a decision tree. In the decision tree, input is entered at the top (root) and then down (leaf) to determine the data, including the type of class order [18]. The random forest is a classifier consisting of a set of structured tree classifiers where each tree issues a sound unit for the most popular class in the x input. Random Forest consists of a collection of decision trees, where a collection of decision trees is used to classify data into a class [21]. The random forest is one of several machine learning algorithms that can be used to perform sentiment analysis. Random forest is a machine learning algorithm that is based on a decision tree algorithm. The basis of the random forest algorithm is a combination of several decision tree algorithms [31]. The process of merging is called ensemble learning using the vote method [32]. The method of vote means selecting the results issued by each decision tree algorithm the most.

The process of inputting variables from several decision tree algorithms is done randomly. Random process is carried out to make each decision tree algorithm disconnected and affect each other [33]. The randomization process uses the bagging method (bootstrap aggregating) which allows taking random input variables and replacing them in the modeling process of each decision tree algorithm.

For example, the random forest input variable is in the form of a list of 4,3,2,8,3,1,8 then when the bagging variable is carried out it is taken from being 2,3,3,8,1,6,4. There is a duplication of the variable number 3 taken. This is necessary to make each decision tree algorithm not connected to one another and influence one another. The randomization process also applies to the features used in the random forest algorithm. For example, when there is feature 1, feature 2 and feature 3. In one decision tree algorithm it only uses feature 2 and feature 3, then in other decision tree algorithms it uses feature 1 and feature 2 and so on.

Another function of randomization is to avoid errors and calculation errors resulting from homogeneous data. The final result of the random forest algorithm is the calculation of the dominant vote of all decision tree algorithms used [34]. The random forest algorithm calculation process is similar to human judgment, in this case it is from several decision tree algorithms with the aim of avoiding a calculation error of one individual decision tree algorithm.

2.2.10. Sentiment Analysis

Sentiment analysis is a natural language processing method whose purpose is to determine the emotional sentiment of the analyzed text [35]. Sentiment analysis works systematically to identify, extract, and study subjective conditions and information. Sentiment analysis is widely applied to consumer opinion analysis, reviews, survey responses, and social media. The main task of sentiment analysis is to classify the polarity of information in documents both sentences and words [36]. Document polarity is grouped into positive, negative, and neutral.

Sentiment analysis is a text analysis process seen from the perspective of the polarity of the sentiment it has. Each text has a specific meaning and purpose following the source of writing which shows the subjectivity of each author [29]. Opinions on subjectivity affect the judgment results of the person or tool that processes and reads the text. Emotional polarity arises from experiences gained and then felt and then poured out in writing.

The display of emotional polarity contained in the text can be found in the choice of words used such as "disappointed" which means that the text writer has negative emotional polarity. The display of emotional polarity is also shown in the use of text emojis [37]. The use of text emoji has a meaning as an interpretation that shows the author's emotions in the text, for example, the emoji “😊” shows a smiling face so that the text implies positive emoji polarity. Sentiment analysis aims to find the emotional polarity contained in the text to find out the true meaning of the emotional experience felt by the author of the text so that certain insights can be found that can be used in many aspects such as improving the quality of corporate services, natural disaster events, important social events and community response to government policy [38] - [42].

3. Result and Discussion

3.1. Tweet Analysis

The sentiment analysis process begins by collecting tweets from Twitter using the GetOldTweet library using the Python programming language as many as 23,777 tweets. The tweet retrieval and filter criteria are based on the location of all provincial capitals in Malaysia with a radius of 50 miles from the coordinates of the provincial capitals. The time-based filtering of tweet data was carried out in the period from December 1, 2019, when the news about Covid-19 was released to June 17, 2020, which was the time the research was carried out.

The keywords used to filter tweet data are based on words related to the economy during the Covid-19 pandemic outbreak. The words used in this research are "money", "profession", "fired", "muflis", "covid19", "corona", "pandemic", "profit", "loss", "entrepreneur", " buy "," commerce "," ringgit "," business "and" workers ". Data that has been successfully collected through other pre-processing stages such as case folding, data cleaning, lemmatizing, and remove stop words.

The cleaned data enters the labeling sentiment stage using Vader sentiment polarity detection. The results of the labeling stage are used as training data on the random forest modeling algorithm.

The process of testing sentiment analysis is carried out using the random forest machine-learning algorithm. Random forest is used in this research because it is an ensemble learning from a collection of individual trees whose working system uses the voting method. The purpose of the ensemble learning forest is to avoid calculation errors of each individual tree calculation and to increase the output calculation quantity of the algorithm. The analysis process uses the sci-kit-learn library in the python programming language.

The extraction feature used to test and maximize the output quality of the random forest algorithm calculation is TF-IDF (Term Frequency - Inverse Document Frequency) and Count Vectorizer. The selection feature is supported by word splitting using n-grams from 1 to 3 grams. the modeling process is coupled with cross-validation to evaluate the random forest model. The modeling results are shown in the following language in tabular form. The FE attribute means the extraction feature, the N-Gram attribute means the number of grams used, the AC attribute means the level of accuracy, and the CV attribute means the level of cross-validation.

3.2. Sentiment Analysis of each Provincial Capital in Malaysia

Details of the data collected from tweets are shown in the following subsections according to provincial capitals in Malaysia. The attribute “Net Tweets” is the result of pre-processing data and the “Sent” attribute is the result of labeling the polarity of the tweet sentiment using vader sentiment polarity detection. The attribute "Sent" has 3 values, namely -1 means negative, 0 means neutral and 1 means positive.

3.2.1. Alor Setar

The results of the pre-processing and labeling stages on the criteria for the capital "Alor Setar" and the coordinates "6.4910077, 100.1823724" are shown in Table 1.

Table 1. The result of the pre-processing and labelling stages of the capital tweet “Aloe Setar”

Clean Tweets	Sent
siapa yg cakap kalau pm yg memerintah mempengaruhi belajar dulu sembang ni rugi jugak sebab penjualan yg mempengaruhi bila mcm ni kurang	1
kesempatan umur jom ambik ambil sekurang kurangnya pelan perlindungan cover gaji musibah tak tersangka	-1
gembar gemburkan mala nak join bisn aku ni melayu dengki pemala sentiasa flex sijil hal lhdn takkan cari aih	-1
padang tanaman bambara groundnut ataupun nama tempatannya berpotensi diketengahkan tanaman kontan pengusaha mangga harumani menjana musim harumani perlispermai perli	1
aq hargai hormat dkt lelaki yg rajin sanggup balik lewat demi nk cari cari demi nk support keluarga	0

The total number of sentiments on the criteria for the capital city "Alor Setar" is shown in Figure 2.

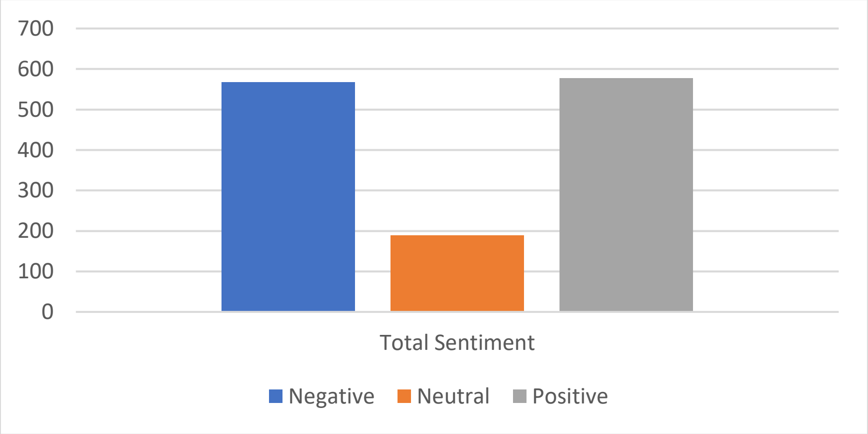

Figure 2. Total sentiment polarity in the capital “Alor Setar”

The results of the sentiment in the capital city of Alor Setar show a not too far difference between positive and negative totals, which is only around 9. Neutral totals tend to have a large selection between positive and negative, the difference is 379 with negative and a total of 388 with positive. Total positive in the capital "Alor Setar" still shows a dominant positive.

In the capital "Alor Setar", the level of accuracy obtained is still 50%, with the highest level being 53.7% and the lowest being 50.7%. The results of the modeling stage are shown in Table 2.

Table 2. Result of modeling in the capital “Alor Setar”.

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,522	0,999	0,617
Count Vectorizer	2	0,537	0,999	0,615
Count Vectorizer	3	0,537	0,999	0,603
TF-IDF	1	0,522	0,999	0,604
TF-IDF	2	0,537	0,999	0,592
TF-IDF	3	0,507	0,999	0,586

3.2.2 George Town

The results of the pre-processing and labeling stages on the criteria for the capital city "George Town" and the coordinates "5.4059643, 100.2743269" are shown in table 3.

Table 3. The results of the pre-processing and labelling stages of the capital tweet "George Town"

Clean Tweets	Sent
journey south tonight may trip run smoothli opshantarpulang usm welead covid n	0
kedah menafikan berita tular jangkitan covid melibatkan kanak kanak kampung huma padang hang penuh berikut	-1
lock covid fightcovid penang penanglawancovid stayathomechalleng staythefhom penang coronaviru staysafestayhom	0
birthday girl turn covid what good babe hahaha thank youu wish	1
journey kuch pleasant flight opshantarpulang usm welead covid n	1

The total sentiment on the criterion for the capital "George Town" is shown in figure 3.

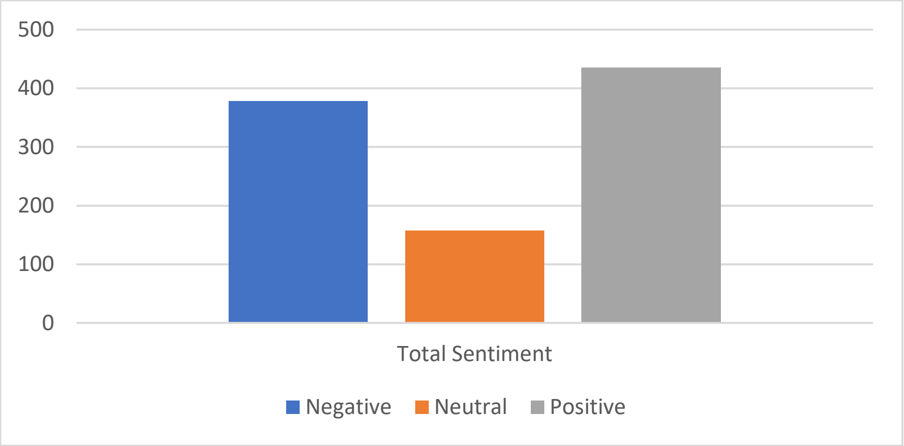

Figure 3. Total sentiment polarity in the capital "George Town"

The sentiment results in the capital city of Gerge Town show quite a difference with a difference of 57 where the results of the dominant sentiment are positive. The total number of neutral tweets is still quite far apart, namely 221 from negative and 278 from positive. The results of sentiment processing show that in the capital "George Town" it is still dominantly positive.

In the capital "George Town" the accuracy rate obtained can reach more than 60% with the highest level of 60.4% and the lowest level of 47.9%. The results of the modeling stage are shown in table 4.

Table 4. Results of modeling in the capital "George Town"

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,604	0,999	0,533
Count Vectorizer	2	0,583	0,999	0,513
Count Vectorizer	3	0,583	0,999	0,535
TF-IDF	1	0,583	0,999	0,512
TF-IDF	2	0,563	0,999	0,510
TF-IDF	3	0,479	0,999	0,490

3.2.3. Ipoh

The results of the pre-processing and labeling stages on the criteria for the capital "Ipoh" and the coordinates "1.5450255, 103.6395867" are shown in table 5.

Table 5. The results of the pre-processing and labeling stages of the capital tweet “Ipoh”

Clean Tweets	Sent
told one want read corona content right wrote anoth corona stori even know know anoth peopl tell want corona content might collect end pandem	0
honestli care corona stat anymor dedic second say omg k move right along	1
new medicin cure corona viru releas medicin name mvi avail mvi moodikittu veetuliy iru day	0
lunch gener host good discuss corona situat upcom ifalpa confer	1
realli hope lanka take necessari measur corona prevent alway better cure lka	1

The total sentiment on the criteria for capitals “Ipoh” shown in the figure 4.

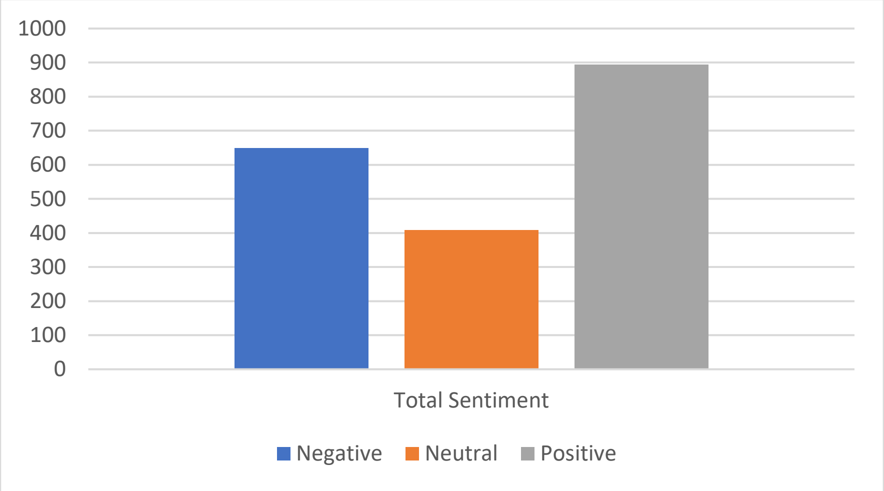

Figure 4. Total sentiment polarity in the capital “Ipoh”

The results of sentiment processing in the capital "Ipoh" showed a dominant positive where the difference of total was 485 with neutral and 244 with negative.
In the capital "Ipoh" the accuracy rate is still below 60% with the highest level being 58.2% and the lowest level being 49%. The results of the modeling stages are shown in the table 6.

Table 6. Results of modeling in the capital “Ipoh”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,582	0,999	0,577
Count Vectorizer	2	0,571	0,999	0,543
Count Vectorizer	3	0,490	0,999	0,521
TF-IDF	1	0,561	0,999	0,561
TF-IDF	2	0,571	0,999	0,529
TF-IDF	3	0,510	0,999	0,539

3.2.4. Johar Bahru

The results of the pre-processing and labeling stages on the capital criteria “Johar Bahru” and coordinates “4.6101207, 101.0215671” shown in the table 7.

Table 7. The results of the pre-processing and labeling stages of the capital tweet “Johar Bahru”

Clean Tweets	Sent
berapa je takkan korang harap pemaju ambil sikit je sebab tu tak murah dibina gaji ni mesti cari tambahan gaji rm beli rm bayar hampir rm k gaji keperluan hidup macam	1
halau asal tandatangan pelarian new zealand canada etc lagipun kalau labur duit pelarian ni kene tingkatkan jangan terkejut cukai nanti tu	0
point anda sdh kasih kesimpulan soal besaran rasio hutang pdb bo anda cm menghitung pembagian umr tertinggi umr jkt kebencian kebodohan yg sdg anda pamerkan	-1
giler duit tuh dia bawak lari tak rasa sebersalah telan duit haram muka lawa tapi sayang takhal	-1
berita sekitar johor anda ingin menjana tambahan kedai anda diperlukan margin	1

The total sentiment on the criteria for capitals “Johar Bahru” shown in the figure 5.

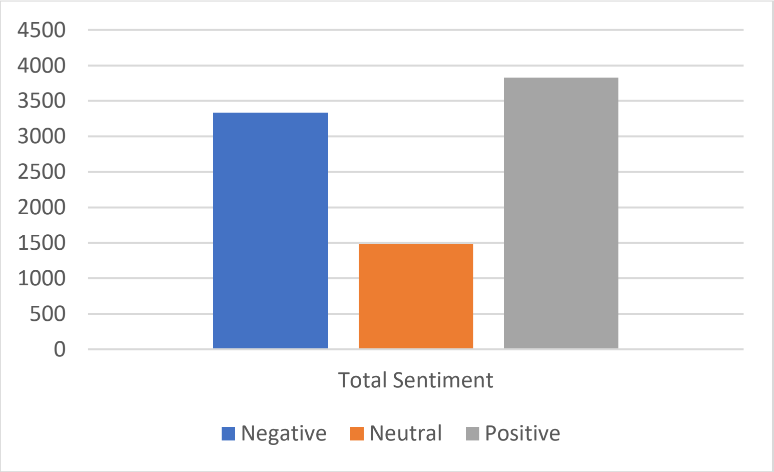

Figure 5. Total sentiment polarityin the capitals “Johar Bahru”

The results of the sentiment in the capital "Johar Bahru" show a neutral total which has a fairly high difference between negative and positive totals. The total difference between positive and negative is a total of 500 which indicates the dominant positive sentiment results.

In the capital city "Johar Bahru" the accuracy rate is still below 60% with the highest level at 59.2% and the lowest level at 49%. The results of the modeling stages are shown in table 8.

Table 8. Results of modeling in the capital “Johar Bahru”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,592	0,999	0,577
Count Vectorizer	2	0,571	0,999	0,543
Count Vectorizer	3	0,490	0,999	0,521
TF-IDF	1	0,561	0,999	0,561
TF-IDF	2	0,571	0,999	0,529
TF-IDF	3	0,510	0,999	0,539

3.2.5. Kangar

The results of the pre-processing and labeling stages on the capital criteria "Kangar" and coordinates "6.4910077, 100.1823724" shown in the table 9.

Table 9. The results of the pre=processing and labeling stages of the capital tweet "Kangar"

Clean Tweets	Sent
nung sumiklab ang u iran kabang kaba ksi sa iraq nung mga ora na yon ta ngayon dun nman patungo sa kung saan sikat tong corona hayop nato sa mag stay pa ng mahigit month dahil drydock	0
easier without mask mean corona though	1
realli cool plan cancel corona	-1
annoy corona viru pleas make stop dont need anoth delusion talentless copycat hoe trynna get everyon attent	-1
corona come uum guy let welcom	1

The total sentiment on the criteria for capitals "Kangar" shown in the figure 6.

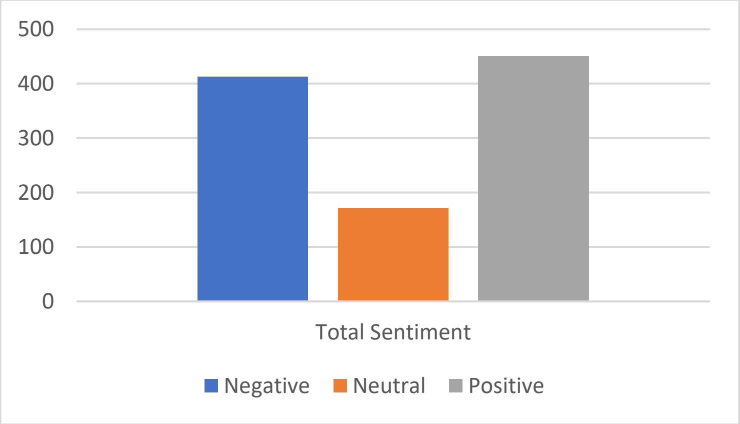

Figure 6. Total sentiment polarity in the capital "Kangar"

The result of sentiment processing in the capital "Kangar" shows a neutral total which is slightly less compared to a negative and positive total. Positive sentiment is still dominant with a slight difference of 38 with negative sentiment.

In the capital "Kangar" the accuracy rate is close to 70% with the highest rate of 69.2% and the lowest level of 61.5%. The results of the modeling stages are shown in table 10.

Table 10. Results of modeling in the capital "Kangar"

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,692	1,000	0,579
Count Vectorizer	2	0,615	1,000	0,568
Count Vectorizer	3	0,654	1,000	0,551
TF-IDF	1	0,615	1,000	0,566
TF-IDF	2	0,692	1,000	0,558
TF-IDF	3	0,654	1,000	0,547

3.2.6 Kota Baharu

The results of the pre-processing and labeling stages on the capital criteria “Kota Bharu” and coordinates “6.1248062, 102.2456277” shown in the table 11.

Table 11. The results of the pre-processing and labeling stages of the capital tweet “Kota Baharu”

Clean Tweets	Sent
ni small scale outbreak pandem consequ would sever doctor pun takda nak tolong	-1
diberi kesempatan bersyukur merasai nikmat ramadhan syawal berbeza pandemik covid terkandung hikmah selamat maaf zahir amp batin staysaf normabaru jagakebersihan aidilfitri selamathariraya syawal covid maafzahirdanbatin	1
stay home comfort without think abt incom work studi frontlin motiv job realli amaz work overcom pandem thankyou thankyou everyon stay home	1
hmm seram nk bukak twitter sbnrnye ni sbb tl penuh dgn hot dkt u tu ujian berat buat u dgn lead pandem tunjuk perasaan dgn yg tak pernah berkesudahan bertahun lamanya may god bless	-1
ni dah kira bodoh sain dah ni kalau pandem pun dia lemah kesian	-1

The total sentiment on the criteria for capitals “Kota Bharu” shown in the figure 7.

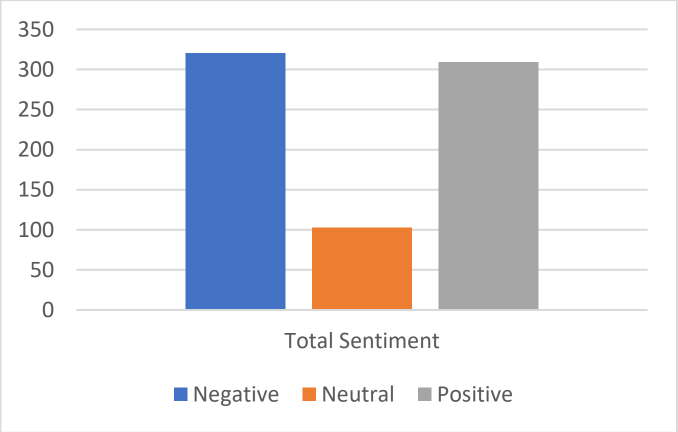

Figure 7. Total sentiment polarity in the capital “Kota Baharu”

The results of processing sentiment in the capital "Kota Bharu" show the same neutral status as the other capitals. But the difference is, the dominant result is a negative sentiment with a difference of 11 from positive sentiment.

In the capital city "Kota Bharu" the accuracy rate is still below 60% with the highest rate of 56.8% and the lowest level of 48.6%. The results of the modeling stages are shown in table 12.

Table 12. Results of modeling in the capital “Kota Baharu”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,541	1,000	0,602
Count Vectorizer	2	0,541	1,000	0,575
Count Vectorizer	3	0,514	1,000	0,562
TF-IDF	1	0,541	1,000	0,569
TF-IDF	2	0,568	1,000	0,541
TF-IDF	3	0,486	1,000	0,538

3.2.7. Kota Kinabalu

The results of the pre-processing and labeling stages on the capital criteria “Kota Kinabalu” and coordinates “5.9961432, 116.0254819” shown in the table 13.

Table 13. The results of the pre-processing and labeling stages of the capital tweet “Kota Kinabalu”

Clean Tweets	Sent
even though price gasolin drop still get hous youjour jrnshahalam covid	0
love natur natur love u youjour jrnshahalam covid	1
behind covid thingi got spend time famili work done home fix clean everi nook cranni hous upgrad cook skill learn imam mekah	1
info wheel tanahmerah haru brtggjwb atsd diri kluarga amp komun kit kawalan kendiri merupakn kunci dlm membendung covid amalkn norma hidup amalkn w amp elak c patuhi sop pkpp kitajagakita kitamestimenang	1
somebodi tri develop vaccin covid home makeshift lab	0

The total sentiment on the criteria for capitals “Kota Kinabalu” shown in the figure 8.

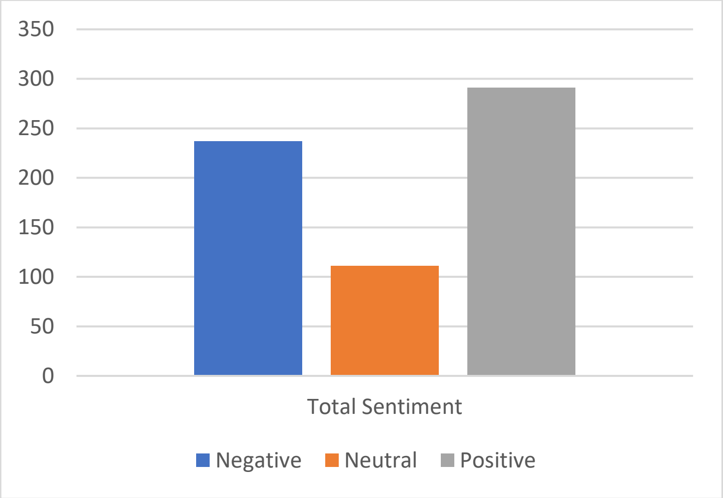

Figure 8. Total sentiment polarity in the capital “Kota Kinabalu”

The results of sentiment processing in the capital city "Kota Kinabalu" show the dominant side of positive sentiment with a difference of 54 negative and 180 neutral.

In the capital city "Kota Kinabalu" the accuracy rate is still below 60%, with the highest level being 56.3% and the lowest level being 40.6%. The results of the modeling stages are shown in table 14.

Table 14. Results of modeling in the capital “Kota Kinabalu”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,469	1,000	0,563
Count Vectorizer	2	0,469	1,000	0,546
Count Vectorizer	3	0,438	1,000	0,510
TF-IDF	1	0,406	1,000	0,528
TF-IDF	2	0,563	1,000	0,513
TF-IDF	3	0,500	1,000	0,494

3.2.8 Kota Melaka

The results of the pre-processing and labeling stages on the capital criteria “Kota Melaka” and coordinates “2.2375488, 102.1814631” shown in the table 15.

Table 15. The results of the pre-processing and labeling stages of the capital tweet "Kota Melaka"

Clean Tweets	Sent
corona name viru wuhan suspect place origin viru ha gitcheww	1
tapi seriu aku kena corona dulu pun demam tekak manja	-1
corona kill peopl alreadi	1
corona viru ah shit go	1
jb pleas la buat world tour corona end	1

The total sentiment on the criteria for capitals "Kota Melaka" shown in the figure 9.

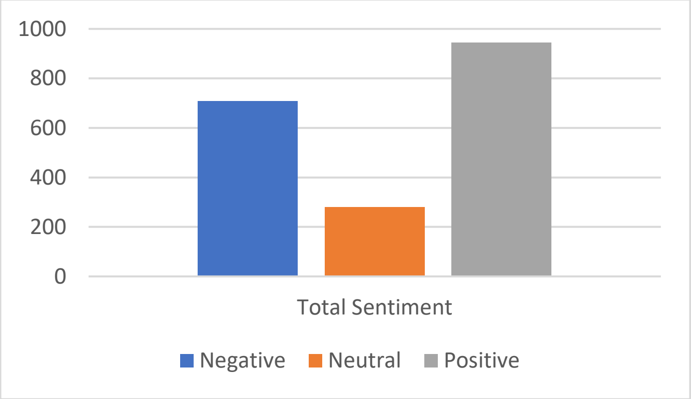

Figure 9. Total sentiment polarity in the capital "Kota Melaka"

The results of sentiment processing in the capital city of "Malacca City" are still dominant compared to neutral and negative sentiment. Negative sentiment has a difference of 238 with positive sentiment and 428 with neutral sentiment.

In the capital city of "Malacca City" the accuracy rate reaches 62.9% with the lowest rate of 52.6%. The results of the modeling stages are shown in the table 16.

Table 16. Results of modeling in the capital "Kota Melaka"

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,598	0,999	0,601
Count Vectorizer	2	0,557	0,999	0,584
Count Vectorizer	3	0,546	0,999	0,562
TF-IDF	1	0,629	0,999	0,590
TF-IDF	2	0,588	0,999	0,581
TF-IDF	3	0,526	0,999	0,584

3.2.9. Kuala Terengganu

The results of the pre-processing and labeling stages on the capital criteria "Kuala Terengganu" and coordinates "5.4845334, 102.7477997" shown in the table 17.

Table 17. The results of the pre-processing and labeling stages of the capital tweet "Kuala Terengganu"

Clean Tweets	Sent
cf close friend chem fun corona fun	1
ni dgn corona sekali energi ni	1
hahahahahahahahaahah cat translat meow meow meow meow meow corona	0
corona fake	1
prepar shit tape avail hostil record soon mean time fun corona	1

The total sentiment on the criteria for capitals “Kuala Terengganu” shown in the figure 10.

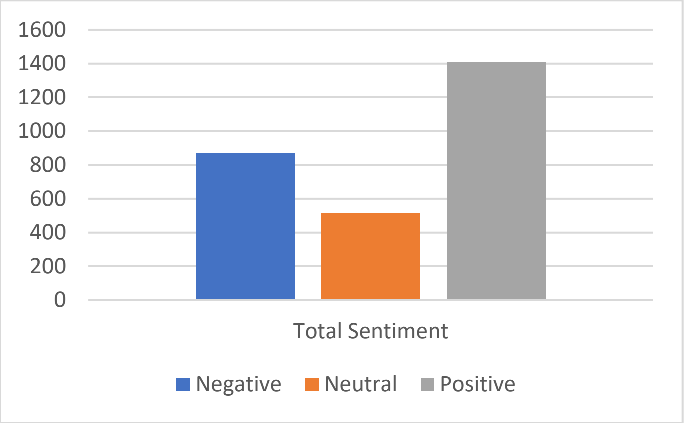

Figure 10. Total sentiment polarity in the capital “Kuala Terengganu”

The results of sentiment processing in the capital city "Kuala Trengganu" show a very strong positive sentiment side compared to the negative sentiment which has a difference of 536 and 894 with the neutral sentiment.
In the capital city of "Kuala Trengganu" the accuracy rate is above 60% with the highest level of 66.4% and the lowest level of 62.9%. The results of the modeling stages are shown in table 18.

Table 18. Results of modeling in the capital “Kuala Terengganu”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,664	0,999	0,617
Count Vectorizer	2	0,643	0,999	0,588
Count Vectorizer	3	0,629	0,999	0,562
TF-IDF	1	0,636	0,999	0,607
TF-IDF	2	0,643	0,999	0,591
TF-IDF	3	0,643	0,999	0,580

3.2.10. Kuantan

The results of the pre-processing and labeling stages on the capital criteria “Kuantan” and coordinates “2.7126609, 101.901953” shown in the table 19.
Table 19. The results of the pre-processing and labeling stages of the capital tweet “Kuantan”

Clean Tweets	Sent
corona corona go away	0
coronavirusfromminorchastis final remind allah command one suprem entir decis maker leader muslim countri imam nasser mohammad al yemeni yamani com showthread php p corona covid corono viru	1
posit corona viru rm give full tank	1
fresh graduat meme next year im gon na employe engin huge compani kadaisil corona vanth enakk planta sapu pannerechiy	1
dearest happi birthday cuti last year got cake tb je mampu sebab corona may god bless good health offer come earlier alreadi gon na work togeth next month hahahaha stay bubbl	1

The total sentiment on the criteria for capitals “Kuantan” shown in the figure 11.

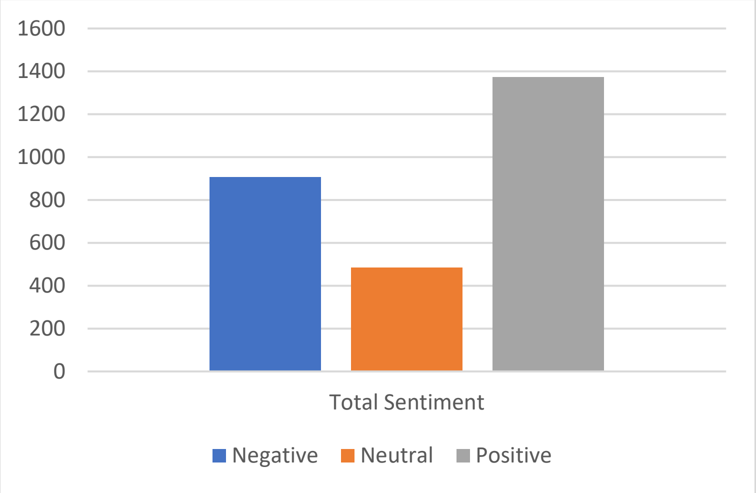

Figure 11. Total sentiment polarity in the capital “Kuantan”

The results of processing sentiment in the capital city "Kuantan" showed a dominant positive side again the same as the capital "Kuala Terengganu" with a total positive sentiment of 1371, the difference of 465 with negative sentiment and the difference of 886 with the neutral sentiment.

In the capital "Kuantan" the accuracy rate is below 70% with the highest level of 63% and the lowest level of 52.9%. The results of the modeling stages are shown in table 20.

Table 20. Results of modeling in the capital “Kuantan”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,587	1,000	0,625
Count Vectorizer	2	0,630	1,000	0,604
Count Vectorizer	3	0,529	1,000	0,568
TF-IDF	1	0,623	1,000	0,616
TF-IDF	2	0,587	1,000	0,597
TF-IDF	3	0,616	1,000	0,589

3.2.11. Kuching

The results of the pre-processing and labeling stages on the capital criteria “Kuching” and coordinates “1.6185192, 110.1859814” shown in the table 21.

Table 21. The results of the pre-processing and labeling stages of the capital tweet “Kuching”

Clean Tweets	Sent
cari intern covid season ni anoth level stress compani suppos take u suddenli cancel alasan covid	0
bantera pati habi habisan imigresen jim covid herocovid frontlin stayathom staysaf dudukrumah kitajagakita kitamestimenang stori fbid amp id	1
power import watch covid socialdistanc jagajarak	1
polic round ensur public juststayathom pleas la guy covid cyberjaya cyberjaya	1
may th boe ja star war r liveri starwarsfan starwar covid	1

The total sentiment on the criteria for capitals “Kuching” shown in the figure 12.

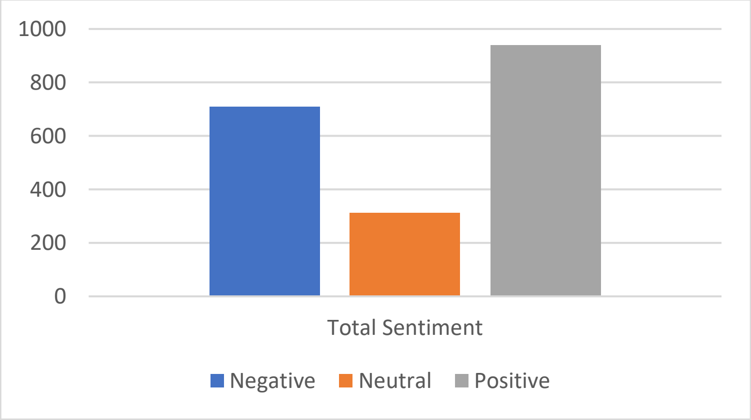

Figure 12. Total sentiment polarity in the capital "Kuching"

The results of sentiment processing in the capital city "Kuching" showed the dominant result in the form of positive sentiment with a difference of 626 with neutral sentiment and 230 with negative sentiment.

In the capital city "Kuching" the accuracy rate is below 70% with the highest rate of 63.3% and the lowest rate of 55.1%. The results of the modeling stages are shown in table 22.

Table 22. Results of modeling in the capital "Kuching"

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,582	0,999	0,594
Count Vectorizer	2	0,551	0,999	0,581
Count Vectorizer	3	0,592	0,999	0,574
TF-IDF	1	0,633	0,999	0,586
TF-IDF	2	0,612	0,999	0,574
TF-IDF	3	0,612	0,999	0,574

3.2.12. Seremban

The results of the pre-processing and labeling stages on the capital criteria "Seremban" and coordinates "2.712322, 101.9019532" shown in the table 23.

Table 23. The results of the pre-processing and labeling stages of the capital tweet "Seremban"

Clean Tweets	Sent
nowday call boo corona real	0
govern learn much new zealand punya govern takkan nak pandang sinc pon fail handl corona issu respect womenpow	1
hope wont get inflect corona viru	1
corona time wat call ci al cing	0
corona kill	1

The total sentiment on the criteria for capitals "Seremban" shown in the figure 13.

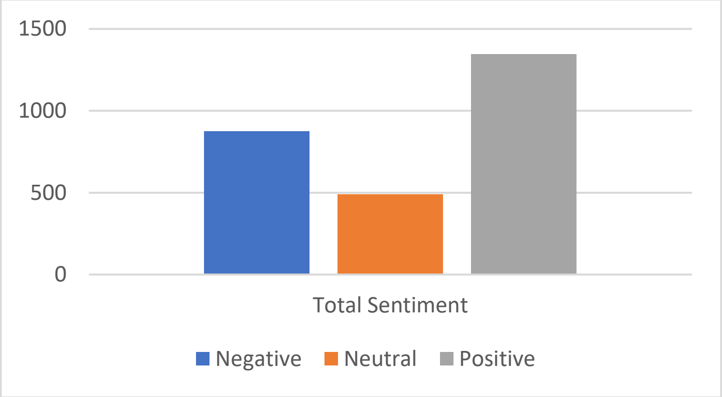

Figure 13. Total sentiment polarity in the capital “Seremban”

The results of the processing of sentiment in the capital "Seremban" showed the same dominant positive as the previous capital with a difference of 856 with neutral sentiment and 470 with negative sentiment.

In the capital city "Seremban" the accuracy rate is below 70% with the highest rate of 64% and the lowest level of 58.8%. The results of the modeling stages are shown in table 24.

Table 24. Results of modeling in the capital “Seremban”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,625	1,000	0,627
Count Vectorizer	2	0,640	1,000	0,599
Count Vectorizer	3	0,588	1,000	0,570
TF-IDF	1	0,632	1,000	0,612
TF-IDF	2	0,618	1,000	0,584
TF-IDF	3	0,603	1,000	0,580

3.2.13 Shah Alam

The results of the pre-processing and labeling stages on the capital criteria “Shah Alam” and coordinates “3.0909411, 101.3764259” shown in the table 25.

Table 25. The results of the pre-processing and labeling stages of the capital tweet “Shah Alam”

Clean Tweets	Sent
hope corona could end ramadan go tarawih bazar iftaar famili friend peac inshallah bazar ramadhan tanjung karang	1
start joke whole world shake miss corona	0
chang govern th gener elect noth corona viru someon think epidem pandem occur chang govern mental coronarviru	0
corona time	0
put corona viru asid know new viru call hantaviru	0

The total sentiment on the criteria for capitals “Shah Alam” shown in the figure 14.

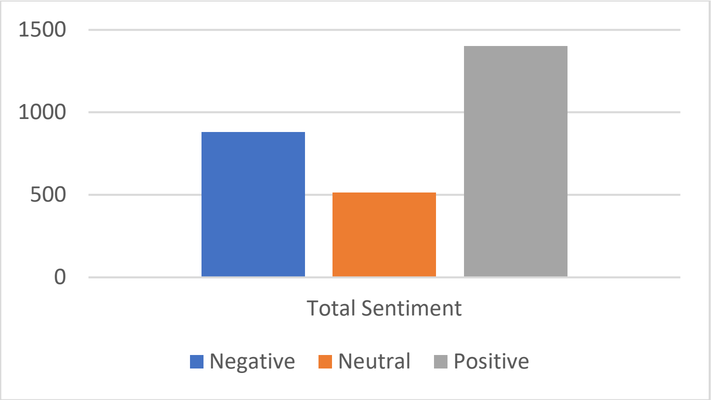

Figure 14. Total sentiment polarity in the capital “Shah Alam”

The results of sentiment processing in the capital "Shah Alam" show a dominant sentiment, namely positive compared to neutral and negative. The difference between positive and negative is 519 and the difference between motive and neutral is 888.

In the capital "Shah Alam" the average accuracy rate is below 60% with the highest level being 60.7% and the lowest level being 52.9%. The results of the modeling stages are shown in table 26.

Table 26. Results of modeling in the capital “Shah Alam”

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,607	1,000	0,610
Count Vectorizer	2	0,571	1,000	0,596
Count Vectorizer	3	0,571	1,000	0,573
TF-IDF	1	0,593	1,000	0,612
TF-IDF	2	0,579	1,000	0,577
TF-IDF	3	0,529	1,000	0,578

3.2.14. Sentiment Analysis in Malaysia

Table 27. Modeling result from all data

FE	N-Gram	AC	CV Train	CV Test
Count Vectorizer	1	0,927	0,999	0,919
Count Vectorizer	2	0,928	0,999	0,915
Count Vectorizer	3	0,929	0,999	0,910
TF-IDF	1	0,935	0,999	0,917
TF-IDF	2	0,934	0,999	0,913
TF-IDF	3	0,933	0,999	0,912

Table 27 shows that the highest accuracy of the random forest algorithm is on the 4th line which shows the use of TF-IDF and 3 grams of data tokens. The accuracy obtained is worth 0.935 with a difference of 0.001 in the 5th row and a difference of 0.002 in the 6th row. The lowest yield is shown in row 1 using a count vectorizer and 1 gram with a value of 0.927.

From table 27 it can be concluded that the calculation process using the count vectorizer is lower than the TF-IDF extraction feature. The more total grams also affects the two extraction features, but there is a difference where in the count vectorizer the more grams the higher the result and on TF-IDF the results will be even lower. The difference between each additional gram is the same, which is 0.001.

4. Conclusions

The Covid-19 pandemic has reached all regions of the world. Including Malaysia. When the Covid-19 outbreak occurs, people are required to comply with government regulations, one of which is not to carry out activities in crowds and public places and to always be at home and always maintain health protocols. Community activities that are usually found in the public environment suddenly stop, which of course affects many areas of personal life and the lives of many people. One of the fields in the field of economics. In the field of economy, in ordinary practice, it involves the interaction of many people, such as direct buying and selling and workers in production. This study analyzes the impact of the Covid-19 pandemic outbreak in the economic sector using social media Twitter through sentiment analysis to determine the polarity of public sentiment. The criteria used in this study are based on the entire location of the provincial capital in Malaysia. Apart from the distance criteria, keywords related to the economy and covid-19 were also determined. The results of research from 23777 tweets that were collected and through the labeling process using Vader sentiment polarity detection showed that public sentiment tended to be positive with a total of 11323 tweets, 4105 neutral and 8349 negative. The modeling process uses a random forest algorithm with the addition of the TFIDF extraction feature and counts vectorizer plus the use of N-Gram. The processing results show that there is a difference in the calculation process that occurs between the use of TFIDF and the count vectorizer. In the TFIDF process, the results tend to increase if the total grams are less which is inversely proportional to the use of the count vectorizer, which tends to increase with the addition of Total grams. The highest accuracy obtained is several 0.935 when using TFIDF with 1 gram and the lowest is 0.927 when using a count vectorizer and 1 gram. In future research, we suggest combining the selection feature and other different extraction features to obtain a high degree of accuracy. We also recommend using other machine learning algorithms to test different levels of accuracy from the use of random forest in this study.

References

[1] R. Korolov *et al.*, “On predicting social unrest using social media,” *Proc. 2016 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2016*, 2016, pp. 89–95, doi: 10.1109/ASONAM.2016.7752218.

[2] N. F. C. Mat, H. A. Edinur, M. K. A. A. Razab, and S. Safuan, “A single mass gathering resulted in massive transmission of COVID-19 infections in Malaysia with further international spread,” *J. Travel Med.*, Vol. 27, No. 3, 2020, pp. 1–4, doi: 10.1093/jtm/taaa059.

[3] M. W. Hasanat, A. Hoque, F. A. Shikha, M. Anwar, A. B. A. Hamid, and H. H. Tat, “The Impact of Coronavirus (Covid-19) on E-Business in Malaysia,” *Asian J. Multidiscip. Stud.*, Vol. 3, No. 1, 2020, pp. 85–90.

[4] V. Balakrishnan, S. Khan, and H. R. Arabnia, “Improving cyberbullying detection using Twitter users’ psychological features and machine learning,” *Comput. Secur.*, Vol. 90, 2020, p. 101710, doi: 10.1016/j.cose.2019.101710.

[5] S. Vashishtha and S. Susan, “Fuzzy rule based unsupervised sentiment analysis from social media posts,” *Expert Syst. Appl.*, Vol. 138, 2019, doi: 10.1016/j.eswa.2019.112834.

[6] V. A. and S. S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” *Int. J. Comput. Appl.*, Vol. 139, No. 11, 2016, pp. 5–15, doi: 10.5120/ijca2016908625.

[7] N. Öztürk and S. Ayvaz, “Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis,” *Telemat. Informatics*, Vol. 35, No. 1, 2018, pp. 136–147, doi: 10.1016/j.tele.2017.10.006.

[8] S. Al-khalifa, I. Aljarah, and M. A. M. Abushariah, “Hate Speech Classification in Arabic Tweets,” *J. Theor. Appl. Inf. Technol.*, Vol. 98, No. 11, 2020, pp. 1816–1831.

[9] S. Das, R. K. Behera, M. Kumar, and S. K. Rath, “Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction,” *Procedia Comput. Sci.*, Vol. 132, No. Iccids, 2018, pp. 956–964, doi: 10.1016/j.procs.2018.05.111.

[10] E. Kušen and M. Strembeck, “Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections,” *Online Soc. Networks Media*, Vol. 5, 2018, pp. 37–50, doi: 10.1016/j.osnem.2017.12.002.

[11] D. Liu and L. Lei, “The appeal to political sentiment: An analysis of Donald Trump’s and Hillary Clinton’s speech themes and discourse strategies in the 2016 US presidential election,” *Discourse, Context Media*, Vol. 25, 2018, pp. 143–152, doi: 10.1016/j.dcm.2018.05.001.

[12] H. M. K. Kumar and B. S. Harish, “Sarcasm classification: A novel approach by using Content Based Feature Selection Method,” *Procedia Comput. Sci.*, Vol. 143, 2018, pp. 378–386, doi: 10.1016/j.procs.2018.10.409.

[13] T. Mustaqim, K. Umam, and M. A. Muslim, “Twitter text mining for sentiment analysis on government’s response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm,” 2020, pp. 8–15, doi: 10.1088/1742-6596/1567/3/032024.

[14] S. Elbagir and J. Yang, “Twitter sentiment analysis using natural language toolkit and Vader sentiment,” *Lect. Notes Eng. Comput. Sci.*, Vol. 2239, 2019, pp. 12–16.

[15] C. J. Hutto and E. E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).”,” *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, 2014, [Online]. Available: http://sentic.net/.

[16] S. Al-Natour and O. Turetken, “A comparative assessment of sentiment analysis and star ratings for consumer reviews,” *Int. J. Inf. Manage.*, Vol. 54, No. August 2019, 2020, p. 102132, doi: 10.1016/j.ijinfomgt.2020.102132.

[17] A. R. Alaei, S. Becken, and B. Stantic, “Sentiment Analysis in Tourism: Capitalizing on Big Data,” *J. Travel Res.*, Vol. 58, No. 2, 2019, pp. 175–191, doi: 10.1177/0047287517747753.

[18] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” *R news*, Vol. 2, No. 3, 2002, pp. 18–22, doi: 10.1177/154405910408300516.

[19] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, “Random forest and support vector machine based hybrid approach to sentiment analysis,” *Procedia Comput. Sci.*, Vol. 127, 2018, pp. 511–520, doi: 10.1016/j.procs.2018.01.150.

[20] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, “Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm,” *Procedia Comput. Sci.*, Vol. 161, 2019, pp. 765–772, doi: 10.1016/j.procs.2019.11.181.

[21] H. Parmar, S. Bhanderu, and G. Shah, “Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters,” 2014.

[22] T. Mustaqim, “Analysis of Public Opinion on Religion and Politics in Indonesia using K-Means Clustering and Vader Sentiment Polarity Detection,” in *Proceeding International Conference on Science and Engineering*, 2020, Vol. 3, pp. 749–754.

[23] S. Dutta, J. Ma, and M. De Choudhury, “Measuring the impact of anxiety on online social interactions,” *12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018*, 2018, pp. 584–587.

[24] M. A. Hassonah, R. Al-Sayyed, A. Rodan, A. M. Al-Zoubi, I. Aljarah, and H. Faris, “An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter,” *Knowledge-Based Syst.*, Vol. 192, 2020, p. 105353, doi: 10.1016/j.knosys.2019.105353.

[25] S. Behrendt and A. Schmidt, “The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility,” *J. Bank. Financ.*, Vol. 96, 2018, pp. 355–367, doi: 10.1016/j.jbankfin.2018.09.016.

[26] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, and E. A. Villaseñor, “A case study of Spanish text transformations for twitter sentiment analysis,” *Expert Syst. Appl.*, Vol. 81, 2017, pp. 457–471, doi: 10.1016/j.eswa.2017.03.071.

[27] A. L. Uitdenbogerd, “World cloud: A prototype data choralification of text documents,” *J. New Music Res.*, Vol. 48, No. 3, 2019, pp. 253–263, doi: 10.1080/09298215.2019.1606255.

[28] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” *Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017*, No. Icwsm, 2017, pp. 512–515.

[29] H. Saif, Y. He, M. Fernandez, and H. Alani, “Contextual semantics for sentiment analysis of Twitter,” *Inf. Process. Manag.*, Vol. 52, No. 1, 2016, pp. 5–19, doi: 10.1016/j.ipm.2015.01.005.

[30] M. Ghiassi and S. Lee, “A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach,” *Expert Syst. Appl.*, Vol. 106, 2018, pp. 197–216, doi: 10.1016/j.eswa.2018.04.006.

[31] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, Vol. 7376 LNAI, 2012, pp. 154–168, doi: 10.1007/978-3-642-31537-4_13.

[32] V. Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, “COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation,” *Proc. IEEE Int. Conf. Comput. Vis.*, Vol. 2015 International Conference on Computer Vision, ICCV 2015, 2015, pp. 3253–3261, doi: 10.1109/ICCV.2015.372.

[33] S. Del Río, V. López, J. M. Benítez, and F. Herrera, "On the use of MapReduce for imbalanced big data using Random Forest," *Inf. Sci. (Ny).*, Vol. 285, No. 1, 2014, pp. 112–137, doi: 10.1016/j.ins.2014.03.043.

[34] V. Y. Kullarni and P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," *Int. J. Adv. Comput.*, Vol. 36, No. 1, 2013, pp. 1144–1156.

[35] R. Nimesh, P. Veera Raghava, S. Prince Mary, and B. Bharathi, *A Survey on Opinion Mining and Sentiment Analysis*, Vol. 590, No. 1. 2019.

[36] S. K and F. F, "Survey on aspect-level sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, Vol. 28, No. 3, 2016, pp. 813–830.

[37] E. Gabarron, E. Dorronzoro, O. Rivera-Romero, and R. Wynn, "Diabetes on Twitter: A Sentiment Analysis," *J. Diabetes Sci. Technol.*, Vol. 13, No. 3, 2019, pp. 439–444, doi: 10.1177/1932296818811679.

[38] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Futur. Gener. Comput. Syst.*, Vol. 106, 2020, pp. 92–104, doi: 10.1016/j.future.2020.01.005.

[39] L. Terán and J. Mancera, "Dynamic profiles using sentiment analysis and twitter data for voting advice applications," *Gov. Inf. Q.*, Vol. 36, No. 3, 2019, pp. 520–535, doi: 10.1016/j.giq.2019.03.003.

[40] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of Political Sentiment Orientations on Twitter," *Procedia Comput. Sci.*, Vol. 167, 2020, pp. 1821–1828, doi: 10.1016/j.procs.2020.03.201.

[41] P. Tiwari et al., Sentiment Analysis for Airlines Services Based on Twitter Dataset. Elsevier Inc., 2019.

[42] Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," *Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015*, No. March, 2016, pp. 1318–1325, doi: 10.1109/ICDMW.2015.7.