

**IN  
REVIEW**



**Sentiment Analysis of Public Transportation MRT Jakarta on Twitter After  
2 Years Serving the People of Jakarta**

*Thesis Report*

Muhammad Fauzi Maulana  
41518010003

**DEPARTMENT OF INFORMATICS  
FACULTY OF COMPUTER SCIENCE  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2021**



**Sentiment Analysis of Public Transportation MRT Jakarta on Twitter After  
2 Years Serving the People of Jakarta**

*Thesis Report*

Submitted to Complete Terms  
Completed a Computer Bachelor Degree

Created By:  
Muhammad Fauzi Maulana  
41518010003

DEPARTMENT OF INFORMATICS  
FACULTY OF COMPUTER SCIENCE  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2021

## ORIGINALITY STATEMENT SHEET

The undersigned below:

Student Number : 41518010003

Name : Muhammad Fauzi Maulana

Thesis Title : Sentiment Analysis of Public Transportation MRT Jakarta  
on Twitter After 2 Years Serving the People of Jakarta

I am stating that my Thesis Report is my own and not plagiarism. If it is found in my Thesis Report that there is an element of plagiarism, I am ready to get academic sanctions related to it.

Jakarta, 12 January 2022



Muhammad Fauzi Maulana



UNIVERSITAS  
MERCU BUANA

## THESIS PUBLICATION STATEMENT

As an Universitas Mercu Buana student, I, the undersigned below :

Student Name : Muhammad Fauzi Maulana  
Student Number : 41518010003  
Thesis Title : Sentiment Analysis of Public Transportation MRT  
Jakarta on Twitter After 2 Years Serving the People  
of Jakarta

By giving permission and approval of **Non-exclusive Royalty-Free Right** to Universitas Mercu Buana for my scientific work entitled above along with the available devices (if necessary).

With this Non-exclusive Royalty-Free Right, Universitas Mercu Buana has the right to store, transfer/format, manage in the form of a database, administer, and publish my thesis.

Furthermore, for the sake of science development in the Universitas Mercu Buana environment, I give permission to Researcher in the Research Lab of Computer Science Faculty, Universitas Mercu Buana, to use and develop the existing results of the research of my thesis for the research and publication purpose as long as my name is stated as author/creator and the copyright owner.

Hereby I made this statement in truthfulness.

Jakarta, 12 January 2022

UNIVERSITAS  
MERCU BUANA



Muhammad Fauzi Maulana

## THESIS OUTPUT STATEMENT LETTER

As an Universitas Mercu Buana student, I, the undersigned below :

Student Name : Muhammad Fauzi Maulana  
 Student Number : 41518010003  
 Thesis Title : Sentiment Analysis of Public Transportation MRT Jakarta on Twitter After 2 Years Serving the People of Jakarta

Declared that :

1. My Thesis Output is as follows:

No	Output	Type	Status
1	Scientific Publication	Not Accredited National Journal	Submitted ✓
		Accredited National Journal	
		Not Reputable International Journal	Approved
		Reputable International Journal	
Submitted/Published at :	Journal Name	: International Journal of Interactive Mobile Technologies (iJIM)	
	ISSN	: 1865-7923	
	Journal Link	: <a href="https://online-journals.org/index.php/i-jim">https://online-journals.org/index.php/i-jim</a>	
	Published Journal Link File	:	

2. Willing to complete the entire article publication process starting from submitting, revising the article until it is declared that it can be published in the intended journal.
3. Asked to attach a scanned ID card and a statement letter (see the HKI document attachment), for the purpose of registering HKI if needed.

This statement I made in truth.

Acknowledge  
Thesis Supervisor

Jakarta, 12 January 2022

Dr. Ida Nurhaida, MT



Muhammad Fauzi Maulana


## COMMITTEE APPROVAL SHEET

Student Number : 41518010003  
Student Name : Muhammad Fauzi Maulana  
Thesis Title : Sentiment Analysis of Public Transportation MRT  
Jakarta on Twitter After 2 Years Serving the People  
of Jakarta


This Thesis has been examined and tried as one of the requirements to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.


Jakarta, 25 January 2022

Approved

  
(Ardiansyah, ST, MTI)  
Head of Defense Committee

UNIVERSITAS  
MERCU BUANA

  
(Prastika Indriyanti, S. Kom, MCS)  
Defense Committee 1

  
(Vina Ayumi, M.Kom)  
Defense Committee 2

## COMMITTEE APPROVAL SHEET

Student Number : 41518010003  
Student Name : Muhammad Fauzi Maulana  
Thesis Title : Sentiment Analysis of Public Transportation MRT  
Jakarta on Twitter After 2 Years Serving the People  
of Jakarta

This Thesis has been examined and tried as one of the requirements to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 25 January 2022

Approved



(Ardiansyah, ST, MTI)

UNIVERSITAS  
MERCU BUANA

## COMMITTEE APPROVAL SHEET

Student Number : 41518010003  
Student Name : Muhammad Fauzi Maulana  
Thesis Title : Sentiment Analysis of Public Transportation MRT  
Jakarta on Twitter After 2 Years Serving the People  
of Jakarta

This Thesis has been examined and tried as one of the requirements to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 25 January 2022





## COMMITTEE APPROVAL SHEET

Student Number : 41518010003  
Student Name : Muhammad Fauzi Maulana  
Thesis Title : Sentiment Analysis of Public Transportation MRT  
Jakarta on Twitter After 2 Years Serving the People  
of Jakarta

This Thesis has been examined and tried as one of the requirements to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 25 January 2022

Approved



(Vina Ayumi, M.Kom)

UNIVERSITAS  
MERCU BUANA

## VALIDITY SHEET

Student Number : 41518010003  
Student Name : Muhammad Fauzi Maulana  
Thesis Title : Sentiment Analysis of Public Transportation MRT Jakarta  
on Twitter After 2 Years Serving the People of Jakarta

This Thesis has been examined and tried as one of the requirements to obtain a Bachelor's Degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

Jakarta, 25 January 2022

Approved,



(Dr. Ida Nurhaida, MT)  
Thesis Supervisor

UNIVERSITAS

MERCU BUANA

Acknowledge,



(Wawan Gunawan, S.Kom, MT)  
Informatics Thesis Coordinator



(Emini R. Kaburuan, Ph.D.)  
Head of Informatics Department

## PREFACE

Praise be given to God Almighty, who has given His grace and gifts, so the author can successfully finish this thesis report entitled “*Sentiment Analysis of Public Transportation MRT Jakarta on Twitter After 2 Years Serving the People of Jakarta*” just in time. This thesis is structured to fulfill one of the requirements to obtain a bachelor’s degree in the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.

The author realizes that completing this thesis report got much support and guidance from the closest people. Here the author would like to say thank you to:

1. Dr. Ida Nurhaida, MT. as the Thesis Supervisor who has taken the time and provided guidance and direction in preparing this thesis to completion.
2. Prastika Indriyanti, S. Kom, MCS. as the Academic Advisor who also has provided knowledge.
3. Emil R. Kaburuan, Ph.D. as the Head of Informatics Engineering, Faculty of Computer Science, Universitas Mercu Buana.
4. Anis Cherid, SE, MTI. as the Secretary of Informatics International Class, Faculty of Computer Science, Universitas Mercu Buana.
5. Wawan Gunawan, S. Kom., M.T. as the Thesis Coordinator for the Informatics Engineering Study Program, Faculty of Computer Science, Universitas Mercu Buana.
6. My Parents who always provide supports and prayers so that the author can finish this thesis smoothly.
7. My Friends in Informatics Engineering 2018 who have given support.
8. All parties that the author cannot mention one by one who has helped a lot in the preparation of this thesis.

Finally, the author hopes that this thesis can be useful for readers to increase both knowledge and insight.

Jakarta, 12 January 2022  
Muhammad Fauzi Maulana

## TABLE OF CONTENTS

<b>COVER .....</b>	<b>i</b>
<b>TITLE PAGE .....</b>	<b>i</b>
<b>ORIGINALITY STATEMENT SHEET .....</b>	<b>ii</b>
<b>THESIS PUBLICATION STATEMENT .....</b>	<b>iii</b>
<b>THESIS OUTPUT STATEMENT LETTER .....</b>	<b>iv</b>
<b>COMMITTEE APPROVAL SHEET .....</b>	<b>v</b>
<b>VALIDITY SHEET .....</b>	<b>ix</b>
<b>ABSTRAK .....</b>	<b>x</b>
<b>ABSTRACT .....</b>	<b>xi</b>
<b>PREFACE .....</b>	<b>xii</b>
<b>TABLE OF CONTENTS.....</b>	<b>xiii</b>
<b>JOURNAL TEXT .....</b>	<b>1</b>
<b>WORKING PAPER.....</b>	<b>14</b>
<b>PART 1. LITERATUR REVIEW .....</b>	<b>16</b>
<b>PART 2. ANALYSIS AND DESIGN.....</b>	<b>24</b>
<b>PART 3. SOURCE CODE.....</b>	<b>27</b>
<b>PART 4. DATASET .....</b>	<b>50</b>
<b>PART 5. EXPERIMENT STAGE .....</b>	<b>53</b>
<b>PART 6. RESULTS ALL EXPERIMENTS .....</b>	<b>58</b>
<b>BIBLIOGRAPHY .....</b>	<b>66</b>
<b>ATTACHMENT OF HAKI DOCUMENTS .....</b>	<b>69</b>
<b>ATTACHMENT OF CORRESPONDENCE .....</b>	<b>71</b>
<b>ATTACHMENT OF CV .....</b>	<b>72</b>

## JOURNAL TEXT

# Sentiment Analysis of Public Transportation MRT Jakarta on Twitter After 2 Years Serving the People of Jakarta

<https://doi.org/10.3991/ijxx.vx.ix.xxxx>

Muhammad Fauzi Maulana<sup>(✉)</sup>

Universitas Mercu Buana, Jakarta, Indonesia  
41518010003@student.mercubuana.ac.id

Ida Nurhaida

Universitas Mercu Buana, Jakarta, Indonesia

**Abstract**—The purposes of this paper are to analyze the sentiment of the tweets from Twitter about MRT Jakarta after two years of serving the people of Jakarta, extract meaningful information from the negative tweets for suggestions to the management of MRT Jakarta, and compare the performance of the Naïve Bayes Classifier and Support Vector Machine (SVM) in classifying the sentiment in this study. The Mass Rapid Transit (MRT) Jakarta is a train-based public transportation mode in Jakarta that provides the people with the effectiveness of breaking through traffic. As a public transportation mode, evaluation of the services provided is needed to improve the quality of service that MRT Jakarta offers. To overcome this problem, it is necessary to analyze the people's opinions about the MRT Jakarta using the sentiment analysis approach. The methods applied in this study are Naïve Bayes Classifier and SVM. This study shows that the proposed method with SVM has better performance than the Naïve Bayes Classifier in classifying the sentiment of the tweets with 90% accuracy on average. However, Naïve Bayes Classifier is faster than SVM in execution times needed to classify the sentiment with 5 seconds on average. From data visualization using word cloud and text association information was obtained, MRT Jakarta should improve their services to the people of Jakarta in terms of power resources, employees, payment process, the procedure to use the service, and the MRT Jakarta application.

**Keywords**—Sentiment Analysis, Text Mining, MRT Jakarta, Naïve Bayes Classifier, Support Vector Machine

## 1 Introduction

MRT (Mass Rapid Transit) Jakarta is a train-based public transportation mode in Jakarta, Indonesia, with many advantages. The main benefits of MRT Jakarta are the speed and effectiveness of breaking through traffic. A person can travel from the south of Jakarta to the center of Jakarta for less than 1 hour. Compared to other transportation in Jakarta for covering the same distance, it can take more than 1 hour from the south to the center of Jakarta [1]. MRT Jakarta, also known as *Moda Raya Terpadu* Jakarta, has a long story to be established. The construction process was started on October 10th, 2014, then it was completed and inaugurated on March 24th, 2019. Now, MRT Jakarta has been serving the people of Jakarta for two years. Therefore, it is necessary to know the people's opinions about MRT Jakarta and then extract meaningful information from the negative sentiment that could be suggestions for evaluation in the future for MRT Jakarta itself.

Sentiment analysis is a technique for measuring and analyzing the sentiment of opinion from a group of people on a particular topic. Implementing the sentiment analysis study with the machine-learning-based method needs labeled data to train the model [2]. Using the sentiment analysis approach will assist the MRT Jakarta management in evaluating the performance of their service to the people of Jakarta and improving their service so the MRT Jakarta will be better after two years of serving the people of Jakarta. This study has two main objectives; first, to give suggestions to the management of MRT Jakarta for evaluation purposes to improve their service to the people of Jakarta. Second, to find out which method or algorithm between Naïve Bayes Classifier and Support Vector Machine (SVM) has a better performance in handling this case study to become a reference for future research.

This paper is organized as follows; section two summarizes the related works. Section three explains the proposed method for data collection, labeling, pre-processing, visualization, negation handling, oversampling, TF-IDF vectorizer, Naïve Bayes Classifier, Support Vector Machine (SVM), and 10-Fold Cross-Validation. Section four shows the analysis and discussion of the results. Moreover, the last section is the conclusion of this study.

## 2 Related Work

Twitter is one social media widely used by the public to express their opinions regarding a particular topic [2]. That is why Twitter is a rich data source for opinion mining and sentiment analysis. It can happen because many people use their phones every day, providing a large amount of data [3]. Through user-generated content from social media, opinion mining and sentiment analysis become critical to investigate the provided service to identify feedback towards a particular topic [4]. One of the example applications of this opinion mining and sentiment analysis is the detection of sentiment from public opinions about public transportation. For public transportation, the data coming from Twitter has been used to evaluate public transportation services provided. This activity has previously been carried out to see responses and opinions public on Indian Railway Express [2], Public Transportation Integrated in Jakarta [5], and Indonesian Online Transportation such as Gojek and Grab [6], [7], [8].

Rachman et al. (2020) [5] researched transportation integrated called Jak Lingko in Jakarta, Indonesia; the results showed that the public mentioned positive sentiment about MRT Jakarta. Fiarni et al. (2018) [6] researched Indonesian online transportation services review; the result showed that the proposed system with Naïve Bayes Algorithm could classify user opinion with 90% precision and 70% recall. Then based on research about sentiment analysis on the release of the iPhone using Support Vector Machine conducted by Bourequat and Mourad (2021) [9], the result showed that the classification with the SVM method produces 89.21% accuracy, 92.43% precision, 95.53% recall and 93.95% F1 score.

Extracting information from tweets to find meaningful information (insight) using word cloud and text association has previously been carried out. Aswani et al. (2018) [10] used social media analytics to derive insights from Twitter by finding the most popular words used in the discussions in the tweets using word cloud and choosing the most popular words related to Search Engine Marketing (SEM). Then they used an adjacency matrix to find associations between the related words and

hashtags. Their research said that it is evident from the word/hashtag analysis that popular terms include several advisory words and suggest best practices for implementation. Ragini et al. (2018) [11] used word cloud to extract the needed information. Using word cloud, they found that it is evident that words available in the tweets categorized with the lexicon do not contribute much to identifying the polarity of the disaster-related tweets. Putri and Muhajir. (2021) [12] used word cloud and word association to extract meaningful tweets. Using word cloud, they found some words that were most frequently discussed, and they analyzed the result of word association then described the meaningful information obtained.

### 3 Proposed Method

This section explains in detail the methodology carried out in this study. This study began with collecting opinions data from Twitter and labeling the data. Then the data were cleaned through data preprocessing. Further, Naïve Bayes Classifier and Support Vector Machine (SVM) were implemented for classifying the sentiment. Finally, the algorithms were tested to see the accuracy level of the algorithms. After that, there were also data visualization stages of finding out insight from the data that can be suggestions for the management of MRT Jakarta. The proposed method can be seen in Figure 1.



Fig. 1. The proposed method

#### 3.1 Data Collection

Opinions from Twitter were collected by using Python programming language with help from Twitter API. For tweets data collection, API key, API key secret, access token, and access token secret of Twitter application are needed. To get those needs, we need to login into a Twitter account on twitter.com and create a Twitter app on developer.twitter.com. After that, the Twitter application serves the needs [13]. The Twitter API directly communicates with the Source and Sink. The authentication keys and tokens are established to help communicate over the Twitter server. The source is a Twitter account, and the sink is HDFS (Hadoop Distributed File System), where all the tweets are stored [14]. In crawling the data from Twitter, a Python library called Tweepy was used. Tweepy is an open-source python library that provides an easy and convenient way to access Twitter API using Python programming language. It provides classes and methods that represent Twitter's models and API endpoints. For filtering the tweets data, this study used

keywords there are “mrt jakarta,” “#mrtjakarta,” and “@mrtjakarta.” The tweets data used in this study is only the tweets text from Twitter then added with id and sentiment in the labeling process.

The dataset used is public opinions on MRT Jakarta after celebrating its two-year birthday (March, 24th 2021) in Indonesian. The Data Collection process was started from May 2021 until September 2021. As a result of the Data Collection, there were 2620 records initially before the duplicate records, and some useless or unrelated tweets were removed. After the same records and useless or irrelevant tweets are removed, the total of the data has become 1611 records. Examples of the dataset used are shown in Table 1.

### 3.2 Data Labeling

Data labeling is the process of preparing tagged or labeled data sets for the model. The models learn to recognize repetitive patterns in labeled data. After enough labeled data is processed, the models can identify the same patterns in data that have not been labeled. In this study, the data is being labeled with three types: Positive, Neutral, and Negative manually by the writer. The writer labeled the data by looking up each word in every sentence to find out the exact meaning of the sentences and then decided what the sentiment for it is, positive, Neutral, or Negative.

Table 1. Data Collection Sample

id	tweet	sentiment
1	@FaisalBasri bukannya semua investor begitu ya? bapak Pernah cek proyek MRT jakarta yang dibiayai jepang, semua pengadaan barang nya 99% harus dari jepang??? bahkan made in europe pun harus beli dari agen di jepang	Negative
2	@kiovt if we date, gw ajak lo naik mrt dan keliling jakarta sampe malem	Positive
3	dah lama buangett ga main ke pusat Jakarta, ada mungkin sekitar 3thn. Hari ini ke bund HI sekitarnya pun jadi coba naik MRT	Positive
4	Ahirnya naik MRT di Jakarta juga	Neutral
5	Hari gini mrt isinya anak gaul jakarta doang	Negative
6	@dendy_wahyudi Hai, Kak. MRT Jakarta beroperasi pukul 06.00-21.00 WIB (hari libur) dan untuk kedatangan ratangga selang waktu 10 menit pada hari libur. Kami sarankan untuk tiba 30 menit lebih awal dari jam keberangkatan ya Kak. Terima kasih.	Neutral
7	Harga Tiket MRT Jakarta Terbaru Dari Tiap Stasiun - <a href="https://t.co/V5oWSKTubO">https://t.co/V5oWSKTubO</a> <a href="https://t.co/FWiCSRFBWq">https://t.co/FWiCSRFBWq</a>	Neutral



### 3.3 Data Preprocessing

The data coming out from Twitter contains various unnecessary contents such as URLs, emoticons, special symbols, usernames, hashtags, additional whitespace, etc. These should be removed before processing it so that the sentiments generated are accurate [14]. In this study, the data preprocessing stage is divided into two because some tasks are unnecessary for Data Visualization. In the first data preprocessing there were some tasks were carried out. The result of this first data preprocessing is for Data Visualization. Those tasks are as follows:

1. Data Transformation: Transformed the label to become numerical, Positive (1), Neutral (0), and Negative (-1).
2. Data Cleansing: Removal of non-sentiment contents such as URLs, emoticons, special symbols, usernames, hashtags, additional whitespace, number, etc.
3. Case Folding: Conversion of text into lowercase.
4. Text Normalization: Handling the contraction words in the Indonesian Language. For example, “udh” and “sdh” are have the same meaning, then those are changed into “sudah,” so later in the TF-IDF vectorizer process, it is treated as the same word.
5. Tokenization: Breaks down a set of characters in a text into word units (token), distinguishing certain characters that can be treated as word separators or not. For example, whitespace characters are considered word separators, such as enter, tabulation, and space.
6. Single Character Removal: Removing all the single characters in the data because it has no meaning and is useless.

After the first data preprocessing was finished, the tweet’s text was merged back into a sentence. Then, the first data preprocessing results were saved into a CSV file again. For the classification purpose, the second data preprocessing stage are needed. There are several tasks were performed as follows:

7. Tokenization: The tweet’s text should be in token form to perform the stopword removal and stemming tasks. So, this process is breaking down a set of characters in a text into word units (token).
8. Stopword Removal: The example of stopwords in the Indonesian Language are “yang,” “yaitu,” “wah,” “nah,” “ada,” etc. Those stopwords have no meaning and are useless for sentiment analysis, so they should be removed.
9. Stemming: Removes the suffix from a word and reduces it to its root word. For example, in the Indonesian Language, “mendorong” is a word, and its suffix is “men-” if we remove “men-” from “mendorong,” then we will get a base word or root word which is “dorong.”

### 3.4 Data Visualization

In this study, a word cloud was used to visualize tweets’ most frequently used words. The font size indicates the frequency of results; the more significant the font size is, the more frequently a word is used [15]. And then, the Text Association was used to identify and analyze the word patterns associated with other words to obtain important information from the negative tweets. Data Visualization aims to extract information in the form of topics that are most often discussed. Information that can be considered essential and associations between words that appear simultaneously can strengthen information search [16].

### 3.5 Negation Handling

Negation words are able to change the polarity of the text. It is not only used to express the denial or inverse of the sentences but also in conjunction with the yes or no questions [17]. Negation words in the Indonesian Language are *tidak*, *tanpa*, *tak*, *belum*, and *kurang*. This study found that 293 negation words exist in the dataset. Those are tagged with “not\_” to handle misclassification in predictions of the algorithms.

### 3.6 Oversampling

The imbalanced dataset is one of the real-world classification challenges using machine learning. An imbalanced dataset can reduce classification accuracy and errors in diagnosis [18]. Getting the best accuracy depends not only on the algorithm used; the character factor dataset used has considerable influence. Improving accuracy can be done if the dataset has a balanced class composition [19]. For handling the imbalanced dataset, there is a technique called oversampling. Oversampling is a technique to enlarge samples in minority classes until they are a majority class. The increasing sampling is done by replication the instance of a minority class or adding new instances by a random subset of the minority class [18].

In this study, oversampling was implemented by upsampling or replicating the minority classes in the data using `resample` function from the `sklearn utils` library in Python. The dataset has 18.62% negative sentiments, 33.95% positive sentiments, and 47.42% neutral sentiments, as shown in Figure 2. After the oversampling process, the data becomes balanced, so all the classes have 33.33% represented records in the dataset, as shown in Figure 3.

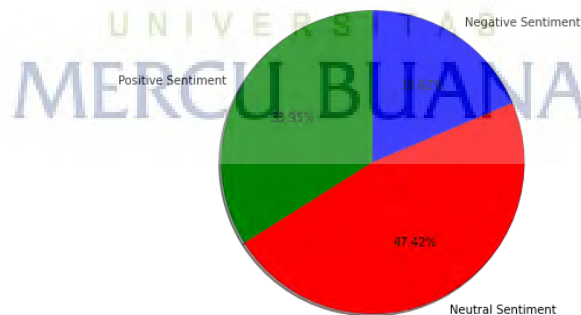


Fig. 2. The dataset before oversampling was performed

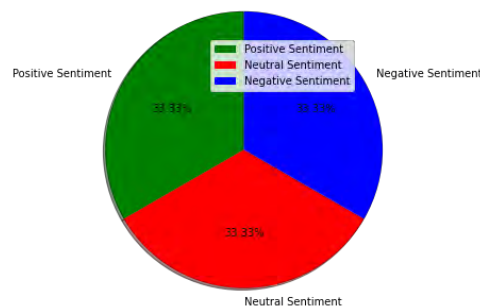


Fig. 3. The dataset after oversampling was performed

### 3.7 TF-IDF Vectorizer

This study used TF-IDF Vectorizer for feature extraction. Feature extraction is essential in text categorization, such as sentiment analysis [20]. The Term Frequency – Inverse Document Frequency (TF-IDF) is a word weighting process where the word will be extracted into the form of feature vectors [21]. This feature extraction process is aimed to find out how important a term is in representing a sentence [22]. This study used TF-IDF Vectorizer from sklearn feature extraction library with n-gram tuple (1,3), which means using unigram, bigram, and trigram.

### 3.8 Naïve Bayes Classifier

Naïve Bayes Classifier is a supervised classification algorithm that uses probability and opportunity approaches. It calculates future probability predictions from data that has been given in the training stage [23]. This algorithm calculates the probability of data belonging to which category using the Bayesian theorem with strong (naïve) independence assumptions between the features in machine learning [24], [25], as we can see in Equation 1. Naïve Bayes Classifier has been studied broadly since the 1950s and remains a standard method for text categorization [25]. Naïve Bayes can also analyze the variables that most influence in the form of probabilities [26]. This study implemented Gaussian Naïve Bayes Classifier and Multinomial Naïve Bayes Classifier to classify the sentiment about MRT Jakarta.

$$P(H_j|x) = \frac{P(x|H_j)P(H_j)}{P(x)} \quad (1)$$

where:

$P(H_j|x)$  = states the probability arises  $H_j$  if known  $x$

$P(x|H_j)$  = The likelihood function of  $H_j$  to  $x$

$P(H_j)$  = Prior probability

$P(x)$  = The evidence

### 3.9 Support Vector Machine (SVM)

Support Vector Machine or SVM is a supervised classification algorithm with a statistical classification approach to maximize the margin between the instances and the separation hyperplane [27]. This algorithm was developed in the 1990s based on the theoretical considerations of Vladimir Vapnik on the development of a statistical theory of learning which is Vapnik-Chervonenkis theory [28].

SVM has a high degree of accuracy in terms of text classification [28]. It uses a separating hyperplane or a decision plane to demarcate decision boundaries among a set of data points classified with different labels. This algorithm finds the hyperplane, which gives the training examples the largest minimum distance (margin) [29]. In this study, we implemented several types of Support Vector Machines. SVM Linear Kernel, SVM RBF Kernel, SVM Polynomial Kernel, and SVM Sigmoid Kernel.

### 3.10 10-Fold Cross-Validation

Evaluation is needed to analyze and measure the accuracy of the results obtained by using 10-Fold Cross-Validation. Cross-Validation is statistical; it can better select a model to predict predictive model test errors [30]. There are two subprocesses in the 10-Fold Cross-Validation, one for training the model and one for testing the model. The dataset is randomly divided into ten subsets because we used ten as the K. The process is repeated or iterated ten times. Ten subsets are used once as the testing data and the rest as the training data [31]. The illustration of how 10-Fold Cross-Validation works can be seen in Figure 4.



Fig. 4. 10-Fold Cross-Validation

## 4 Result Analysis and Discussion

The training and testing data we used in this study were crawled from Twitter with 1611 instances labeled as 764 neutral sentiments, 547 positive sentiments, and 300 negative sentiments, as shown in Figure 5. However, after the oversampling were performed to the minority classes, the data have 2292 instances, with all the classes having 764 instances as the representative, as shown in Figure 6.

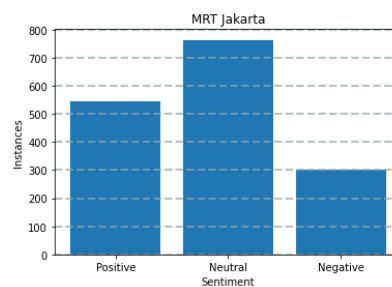
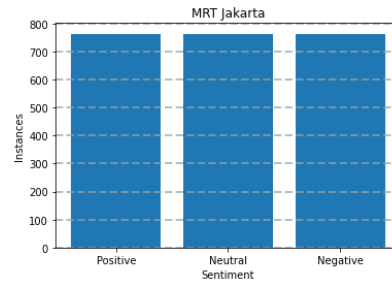


Fig. 5. The original Data



**Fig. 6.** The Data after oversampling stage

In our proposed method, we compare two algorithms there were Naïve Bayes Classifier and Support Vector Machine with 10-Fold Cross-Validation for validating the classification results from both algorithms. The results obtained are shown in Table 2.

**Table 2.** Performance of the Algorithms with 10-Fold Cross-Validation

Algorithm	Accuracy	Execution Times
Gaussian Naïve Bayes (GNB)	84.51%	8.59 Seconds
Multinomial Naïve Bayes (MNB)	85.99%	2.64 Seconds
SVM Linear Kernel	90.57%	9.62 Minutes
SVM RBF Kernel	91.45%	10.82 Minutes
SVM Polynomial Kernel	90.44%	10.11 Minutes
SVM Sigmoid Kernel	88.91%	8.75 Minutes

From Table 2, in terms of the accuracy of the algorithms, we can see that SVM has better performance in classifying the sentiments of the tweets. It has a 90% accuracy average from 4 types of SVM Kernel were performed. SVM RBF Kernel has the highest accuracy among the other algorithms and types. It is because SVM can linearly separate classes by a large hyperplane. It became one of the most powerful classifiers for handling infinite-dimensional feature vectors like text classification, as stated in [26], [27]. However, in terms of the execution times needed, Naïve Bayes Classifier is faster than the SVM. Naïve Bayes only needs 5 seconds on average for classifying the sentiments, while the SVM needs 10 minutes on average for it. Multinomial Naïve Bayes has the fastest execution times among the other algorithms and their types. It is because Naïve Bayes Classifier is just a simple probability-based prediction technique based on the application of Bayes Theorem to assume strong independence. Hence, the way that Naïve Bayes works is more straightforward than SVM, as stated in [24], [25].



Fig. 7. Most Used Words in Negative Sentiment

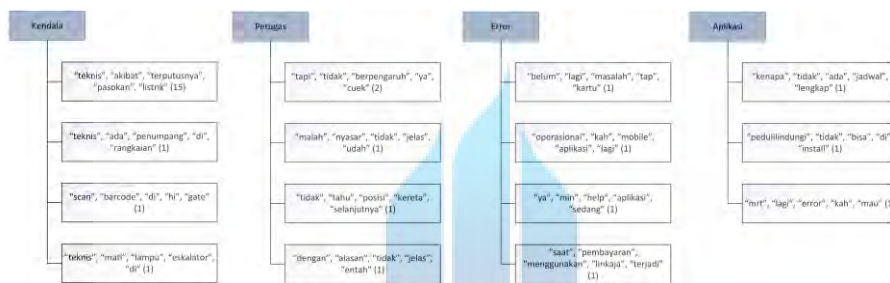


Fig. 8. Text Association in Negative Sentiment

From the Data Visualization using word cloud, as we can see in Figure 7 and text association as we can see in Figure 8, some information from negative sentiments that could be suggestions for the management of MRT Jakarta were obtained. First, about the word “kendala,” which means issue or trouble, there were many MRT Jakarta users who complained about the incident of a power outage which caused the entire MRT Jakarta fleet to stop suddenly on the top lane (flyover) and in the tunnel lane with the passengers in it and also caused all the escalator in the station cannot be used. Moreover, there is an issue with scanning a barcode in the gate for the passenger. Second, about the word “petugas,” which means the employee of MRT Jakarta, some people talked about the employee of MRT Jakarta that did not care about some people who did not apply the health protocol in a station. And then some people complained about the employee who cannot give some information needed for the passengers, such as the train schedule and directions. Third, about the word “error,” some people said they have a problem paying in gate by using the card and linkaja application. Last, about the word “aplikasi,” which means application, people complained that the MRT Jakarta application did not provide them with the schedule of MRT Jakarta. Moreover, some people said they get an error when using the MRT Jakarta application. After that, some people could not install the pedulilindungi application to fulfill the requirement to use MRT Jakarta.

## 5 Conclusion

Based on the study results, it was concluded that the Support Vector Machine (SVM) algorithm has better performance than the Naïve Bayes Classifier in classifying the sentiments of the tweets. The Support Vector Machine (SVM) got a 90% accuracy average from 4 types of SVM kernels were performed. SVM RBF Kernel has the highest accuracy among the other algorithms and their types. SVM can linearly separate classes by a large hyperplane and become one of the most powerful classifiers for handling infinite-dimensional feature vectors like text classification. However, in terms of execution times needed, Naïve Bayes Classifier is faster than SVM. Naïve Bayes only needs 5 seconds on average for classifying the sentiments, while the SVM needs 10 minutes on average for it. Multinomial Naïve Bayes has the fastest execution times among the other algorithms and their types. Naïve Bayes Classifier is just a simple probability-based prediction technique based on the application of Bayes Theorem to assume strong independence, so the way that Naïve Bayes works is more straightforward than SVM. Based on the results of data visualization using word cloud and text association, there were some suggestions for the management of MRT Jakarta to improve their service by solving several issues that occurred in the result. They should improve their services to the people of Jakarta in terms of power resources, employees, payment process, the procedure to use the service, and then the MRT Jakarta Application to make more people change their daily transportation to use MRT Jakarta.

## 6 References

- [1] Rumah, "Panduan Rute MRT Jakarta lengkap Jadwal Dan Harga tiket," *Rumah.com*, 24-Sep-2020. [Online]. Available: <http://www.rumah.com/panduan-properti/rute-mrt-jakarta-33318>. [Accessed: 21-Nov-2021].
- [2] V. C. Rama Krishna, R. Kumar, and Yogita, "Sentiment Analysis of Train Derailment in India: A Case Study from Twitter Data," *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, pp. 230-234, 2019.
- [3] I. Nurhaida, H. Noprisson, V. Ayumi, H. Wei, E. D. Putra, M. Utami, and H. Setiawan, "Implementation of Deep Learning Predictor (LSTM) Algorithm for Human Mobility Prediction," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 14, no. 18, pp. 132-144, 2020.
- [4] M. AbdelFattah, D. Galal, N. Hassan, D. Elzanfaly, and G. Tallent, "A Sentiment Analysis Tool for Determining the Promotional Success of Fashion Images on Instagram," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 11, no. 2, pp. 66-73, 2017.
- [5] F. F. Rachman, R. Nooraeni, and L. Yuliana, "Public Opinion of Transportation Integrated (Jak Lingko), in DKI Jakarta, Indonesia," *Procedia Computer Science*, vol. 179, pp. 696-703, 2021.
- [6] C. Fiarni, H. Maharani, and E. Irawan, "Implementing Rule-based and Naive Bayes Algorithm on Incremental Sentiment Analysis System for Indonesian Online Transportation Services Review," *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 597-602, 2018.
- [7] S. Anastasia and I. Budi, "Twitter Sentiment Analysis of Online Transportation Service Providers," *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 359-365, 2016.
- [8] E. Y. Sari, A. D. Wierfi, and A. Setyanto, "Sentiment Analysis of Customer Satisfaction on Transportation Network Company Using Naive Bayes Classifier," *2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pp. 1-6, 2019.
- [9] W. Bourequat and H. Mourad, "Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine," *International Journal of Advances in Data and Information Systems*, vol. 2, no. 1, pp. 36-44, 2021.

- [10] R. Aswani, A. K. Kar, P. V. Ilavarasan, and Y. K. Dwivedi, "Search engine marketing is not all gold: Insights from Twitter and SEOClerks," *International Journal of Information Management*, vol. 38, pp. 107–116, 2018.
- [11] J. R. Ragini, P. M. R. Anand, and V. Bhaskar, "Big Data Analytics for disaster response and recovery through sentiment analysis," *International Journal of Information Management*, vol. 42, pp. 13–24, 2018.
- [12] Z. N. Putri and M. Muhajir, "Sentiment Analysis of the Large Priest of FPI's Return using Support Vector Machine with Oversampling Method," *Jurnal Riset Informatika*, vol. 4, no. 1, pp. 17–22, 2021.
- [13] M. I. Zul, F. Yulia, and D. Nurmalasari, "Social Media Sentiment Analysis Using K-Means and Naïve Bayes Algorithm," *2018 2nd International Conference on Electrical Engineering and Informatics (ICon EEI)*, pp. 24–29, 2018.
- [14] H. Parveen and S. Pandey, "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm," *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pp. 416–419, 2016.
- [15] P. T. Nguyen, V. Likhitrungsilp, and M. Onishi, "Success Factors for Public-Private Partnership Infrastructure Projects in Vietnam," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 2, pp. 858–865, 2020.
- [16] B. A. Dewanti Putri, A. U. Khasanah, and A. 'Azzam, "Sentiment Analysis on Grab User Reviews Using Support Vector Machine and Maximum Entropy Methods," *2019 International Conference on Information and Communications Technology (ICOIACT)*, pp. 468–473, 2019.
- [17] R. Amalia, M. A. Bijaksana, and D. Darmantoro, "Negation handling in sentiment classification using rule-based adapted from Indonesian language syntactic for Indonesian text in Twitter," *Journal of Physics: Conference Series*, vol. 971, pp. 1–10, 2018.
- [18] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting," *2019 5th International Conference on Science in Information Technology (ICSITech)*, pp. 217–222, 2019.
- [19] M. Hayaty, S. Muthmainah, and S. M. Ghufuran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," *International Journal of Artificial Intelligence Research*, vol. 4, no. 2, pp. 86–94, 2021.
- [20] M. Khalafat, J. S. Alqatawna, R. M. Al-Sayyed, M. Eshtay, and T. Kobbacy, "Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 15, no. 14, pp. 90–110, 2021.
- [21] J. H. Jaman and R. Abdulrohman, "Sentiment Analysis of Customers on Utilizing Online Motorcycle Taxi Service at Twitter with the Support Vector Machine," *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pp. 231–234, 2019.
- [22] I. P. Windasari, F. N. Uzzi, and K. I. Satoto, "Sentiment Analysis on Twitter Posts: An Analysis of Positive or Negative Opinion on Gojek," *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pp. 266–269, 2017.
- [23] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes," *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, pp. 49–54, 2019.
- [24] S. Qaiser, N. Yusoff, F. Kabir Ahmad, and R. Ali, "Sentiment Analysis of Impact of Technology on Employment from Text on Twitter," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 14, no. 07, pp. 88–103, 2020.
- [25] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala, and S. Saxena, "Sentimental Analysis Using Fuzzy and Naive Bayes," *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 945–950, 2017.
- [26] M. Sadikin and F. Alfiandi, "Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 4763–4771, 2018.
- [27] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," *Procedia Computer Science*, vol. 127, pp. 511–520, 2018.
- [28] R. S. A. Corpuz, "Categorizing Natural Language-Based Customer Satisfaction: An Implementation Method Using Support Vector Machine and Long Short-Term Memory Neural Network," *International Journal of Integrated Engineering*, vol. 13, no. 4, pp. 77–91, 2021.
- [29] H. T. Sueno, R. P. Medina, and B. D. Gerardo, "Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3937–3944, 2020.
- [30] Z. R. Tembusai, H. Mawengkang, and M. Zarlis, "K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification," *International Journal of Advances in Data and Information Systems*, vol. 2, no. 1, pp. 1–8, 2021.



- [31] R. Damiaza and D. Fitriana, "Prediction Analysis of Kartu Jakarta Pintar (KJP) Awardees in Vocational High School XYZ Using C4.5 Algorithm," *International Journal of Machine Learning and Computing*, vol. 10, no. 1, pp. 44–50, 2020.

## 7 Authors

**Muhammad Fauzi Maulana** was born in Bekasi, Indonesia in 2001. He is currently pursuing a bachelor's degree in informatics engineering at the Faculty of Computer Science, Universitas Mercu Buana Indonesia. His research interests are data mining, machine learning, and natural language processing. (email: 41518010003@student.mercubuana.ac.id).

**Ida Nurhaida** is a senior lecturer at Universitas Mercu Buana, Faculty of Computer Science, Department of Informatics, Jakarta, Indonesia. Her research interests are Image Processing, Networking, and Deep Learning (email: ida.nurhaida@mercubuana.ac.id)

Article submitted 2019-04-11. Resubmitted 2019-05-27. Final acceptance 2019-05-27. Final version published as submitted by the authors.



## WORKING PAPER

### Summary

This working paper is material for completing the journal article entitled “*Sentiment Analysis of Public Transportation MRT Jakarta on Twitter After 2 Years Serving the People of Jakarta*”. The working paper contains all the research materials of the Thesis which are not published / or included in journal articles. In this paper, the following sections are presented:

1. Literature Review is a section that contains the results of literature studies that have been done related to the experiments carried out. The literature review was conducted on the concept of Sentiment Analysis, Natural Language Processing, Naïve Bayes Classifier, Support Vector Machine, Word Cloud, Text Association, and Imbalance Data.
2. Analysis and Design is a section that consists of an outline and the stages carried out in this research. At this stage is explained that this research is carried out using two scenarios. First, the testing and validation for the classification results are using 10-Fold Cross-Validation. And then, in the second scenario, the data will be split using Train Test Split first and then Training and Testing the model for doing the classification process.
3. Source Code is a section that consists of the code used in this research and the explanations. The Source Code in this study uses Python Programming Language and Google Colaboratory for the IDE. The Python Programming Language's uses in this research are for Tweets Crawling, Data Preprocessing, Data Visualization, Negation Handling, Oversampling, TF-IDF Vectorizer, Implementing the Classification with Naïve Bayes Classifier, and Support Vector Machine (SVM).
4. Dataset is a section that explains what data is used in this research. This section describes the structure of the initial dataset, the treatments performed to the Tweets data from Twitter about MRT Jakarta, and the result.
5. Experimental Stage is a section that contains all experimental stages for completing this research. The stages described in this section include the stages of Data Collection, Data Labeling, Data Preprocessing, Data

Visualization, Oversampling, TF-IDF Vectorizer, Naïve Bayes Classifier, Support Vector Machine (SVM), and 10-Fold Cross-Validation.

6. Results All Experiment is a section consisting of all the experimental results carried out.

