Sentiment Analysis of Public Transportation Services on Twitter Social Media Using the Method Naïve Bayes Classifier

Rima Tamara Aldisa¹, Pandu Maulana², Muhammad Aldinugroho³

¹Faculty of Communication and Information Technology, Universitas Nasional

²Faculty of Computer Science, Universitas Indonesia

³Faculty of Communication and Information Technology, Universitas Nasional

rimatamaraa@gmail.com¹, pandu.maulana@ui.ac.id²,
nugrohoaldi48@gmail.com³

Abstract

Public transportation services in Indonesia, especially Jabodetabek, have used social media, especially Twitter, as a way to improve services. Currently the use of online transportation services is like a need, it is necessary to conduct a sentiment analysis of online transportation to find out how people respond to these online transportation services. This research was made to analyze community responses with data analysis in the form of tweets that filtered with public transportation related keyword then classified into positive and negative classes using the Naïve Bayes Classifier method. Based on the system built, the total sentiment results for the percentage of the occurrence of positive words were 0.507843137 and the sentiment results for the percentage of negative word occurrences were 1.4132493. The results show that the level of negative sentiment from public tweets is greater than the level of positive sentiment.

Keywords: Naïve Bayes Classifier, Public Transportation, Sentiment Analysis, Twitter

1. Introduction

Service (customer service) in general is any activity aimed at providing satisfaction to customers, through service the desires and needs of customers can be fulfilled [1]. For an organization, especially a service organization, customer service is very useful to bind old customers and add new customers. This service can be in the form of complaint management or as a medium of information that the organization can provide to customers. The large and growing use of social media in Indonesia makes organizations use it as a media service. Organizations can use social media to share information and receive complaints from customers. By taking advantage of the advantages of social media, organizations can be more active in managing customer complaints so that customer satisfaction can also be easily achieved. The existence of services on social media by various organizations can also make it easier for customers or prospective customers to get the information they need. However, services on social media can also make it easier for customers to submit reviews and complaints about the organization's services. As the focus of this research is how to take advantage of reviews or complaints on social media from the services provided by this public transportation organization. Indonesian people, especially in Greater Jakarta, use several public transportation services such as Transjakarta, Commuter Line (KRL), Gojek, Grab, etc. These four public transportation organizations have services that are carried out on social media, especially Twitter. Services carried out on Twitter social media look very transparent and effective. With these services on social media, it can be seen the performance of the services offered by each organization with the information submitted by customers. Information can be submitted directly by customers on Twitter social media to review or submit complaints, it can be just a tweet or directly submit it to the organization's social media account. The

advantage of social media Twitter in this regard is that reviews and complaints can be seen by anyone. These reviews and complaints can be used to assess how good or bad the services provided by each organization are. This encourages researchers to research more about the services of public transportation organizations through Twitter.

This study will focus on the service assessment of public transportation services based on reviews or complaints submitted by customers on Twitter social media. The researcher conducted a sentiment analysis on Twitter data by crawling the data on each keyword for the four public transportation services that the researcher adopted. Then do the clarification of the document. There are various techniques for clarifying documents, namely Naive Bayes Classifier, Decision Trees, and Support Vector Machines [2]. And among these the Naive Bayes Classifier technique is the most popular. So that positive and negative values will be generated from reviews of the four public transportation organizations on Twitter social media from each of these public transportation service keywords.

2. Research Methodology

2.1. Sentiment Analysis Framework

Sentiment analysis or sentiment analysis in Indonesian is a technique or method used to identify how a sentiment is expressed using text and how that sentiment can be categorized as positive sentiment or negative sentiment [3]. Almost the same opinion was expressed [4], where sentiment analysis is used to understand comments made by users (internet) and explain how a product or brand is received by them. Meanwhile, sentiment analysis is the process used to determine opinions, emotions and attitudes that are reflected through the text, and are usually classified into negative and positive opinions.

2.2. Naive Bayes Classification

Naïve Bayes Classification is a classification method rooted in Bayes' theorem. The classification method using probability and statistical methods proposed by British scientist Thomas Bayes, which predicts future opportunities based on past experience, is known as Bayes' theorem. The main feature of this Naïve Bayes Classifier is a very strong (naive) assumption of the independence of each condition or event.

According to [6] explains Naïve Bayes for each decision class, calculates the probability provided that the decision class is true, given the object information vector. This algorithm assumes that object attributes are independent. The probabilities involved in producing the final estimate are calculated as the sum of the frequencies from the "master" decision table [6]. The Naive Bayes equation used in this study using Bayes' theorem (Bustami 2013) is [7]:

$$P(H|X) = \frac{P(X|H)}{P(X)} \cdot P(H) \quad (1)$$

Information :

X : Data with unknown class

H : Hypothesis data is a specific class

P(H|X) : Probability of hypothesis H based on condition X (posteriori probability)

P(H) : Hypothesis probability H (prior probability)

P(X|H) : Probability of X based on the conditions on the hypothesis H

P(X) : Probability X

The advantage of using this method is that this method only requires a small amount of training data to determine the parameter estimates needed in the classification process. Because it is assumed to be an independent variable, only the variance of a variable in a class is needed to determine the classification, not the entire covariance matrix. Benefits of Naive Bayes:

- a) Classify text documents such as news texts or academic texts.
- b) As a machine learning method that uses probability
- c) To make medical diagnosis automatically.
- d) Detect or filter spam.

Advantages of Naive Bayes:

- a) Can be used for both quantitative and qualitative data.
- b) Does not require a large amount of data.
- c) No need to do a lot of training data.
- d) If there is a missing value, it can be ignored in the calculation.
- e) The calculation is fast and efficient.
- f) Easy to understand.
- g) Easy to make.
- h) Document classification can be personalized, tailored to the needs of each person.
- i) If used in a programming language, the code is simple.
- j) Can be used for binary or multiclass problem classification.

2.3. Research Instruments

Based on the scope of the research to be studied, the approach of this research is quantitative research which is studied by studying numbers on Twitter social media. The location of this research was conducted on Twitter social media in Indonesia. Then the source of data from this research is from observation and calculation of data that has been collected from social media Twitter.

2.4. Data Collection Procedure

The data collection technique used is the Python programming language. Observations on Twitter social media were carried out using text mining techniques for 5 hours on each keyword. We search for tweets or comments from Twitter users by category as follows:

- 1) 'Gojek'.
- 2) 'Grab'.
- 3) 'Commuter Line'.
- 4) 'Transjakarta'.

2.5. Data Analysis Method

Research conducted using the Python programming language to develop it. The process methodology consists of Data Collection taken from social media Twitter, Data Cleaning, and Naive Bayes Classification. Python is used to display Sentiment Analysis because Python can collect, process, classify, analyze, and loop conditional data based on various types of desired input.

- a) Data Collection.
The raw data is collected from Indonesian-language tweets from all users in Indonesia.
- b) Data Cleaning.
Original data tends to be incomplete, noisy, and inconsistent. Data Cleaning (or data cleaning) will detect missing values, smooth noise while identifying outliers, and correct inconsistencies in the data [8].
- c) Data Integration.
Data mining often requires data integration. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help increase the accuracy and speed of the data mining process.
- d) Data Reduction.
Complex data analysis and mining of large amounts of data can take a long time, making data analysis impractical. Data reduction techniques can be applied to

obtain a reduced representation of a data set whose volume is much smaller, but still maintains the integrity of the data.

e) Data Transformation.

In Data transformation, data is transformed or consolidated into a form suitable for mining.

f) Classification.

At this classification stage, we use a classification with Naive Bayes.

3. Results And Discussion

The data collection technique through Twitter social media is carried out using the Python programming language. The data collected is in the form of tweets which will be processed with positive and negative classification filters. The resulting data has a percentage for each occurrence of the word, both positive and negative. The following is an analysis and discussion of each category using the Naive Bayes calculation method. The Naive Bayes equation used is as follows.

$$P(c|x) = P(x_1|c) P(x_n|c) \dots P(c) \quad (2)$$

$P(c|x)$ = Probability based on condition

$P(x|c)$ = Probability based on hypothetical conditions

$P(c)$ = Probability of the whole hypothesis

The following is the crawled data for the keyword 'Gojek' using the Python programming language.

Table 1. Percentage Of Gojek Users Sentiment

Tweet	Type Tweet		Type Tweet (%)	
	Positif	Negatif	Positif (%)	Negatif (%)
1	[]	[' Expensive']	0	12,5
2	[]	[' Expensive']	0	10
3	[]	['slow']	0	20
4	[]	['slow']	0	25
5	[]	['bad']	0	12,5
6	[]	[Expensive ', 'lazy']	0	33,3333333
7	[]	[' congested ']	0	7,69230769
8	[' steady ']	[]	7,14285714	0
9	[]	[' complicated ']	0	4,76190476
10	[]	['slow']	0	5
11	['good', 'love']	[]	15,3846153	0
12	[]	['long', ' congested ']	0	11,11111111
13	[cheap ']	[]	11,11111111	0
14	[cheap ']	[]	8,33333333	0

Table 2. Results of Gojek Respondents' Sentiments

Type	Positif	Negatif
Probabilitas	0,285714286	0,714285714
Mean	2,997994069	10,13561835

There are 14 tweets with each classification. There are 4 positive tweets with a probability of 0.285714286, and negative tweets with a probability of 0.714285714. The percentage for each type of tweet is 2.997994069% for positive, and 10.13561835% for negative. For classification related to word occurrence data, it is in the following table.

Table 3. Gojek User Data Opinion Factors

Type	word occurrence	Probabilitas	
		Positif	Negatif
Expensive'	3	0	0,18
slow	3	0	0,18
bad	1	0	0,06
lazy	1	0	0,06
congested	2	0	0,12
steady	1	0,06	0
complicated	1	0	0,06
good	1	0,06	0
love	1	0,06	0
long	1	0	0,06
cheap	2	0,12	0
Total	17	0,29	0,71

In table 3, there are occurrences of words with their respective numbers. The probability of the occurrence of a word is the result of each occurrence of the word among the total number of words that appear in the tweet, which is 17. The probability result of the occurrence of a positive word is 0.29 and a negative one is 0.71. The following is the result of calculating the total probability using the Naive Bayes method.

Table 4. Results Of Calculation Of Total Probability Of Gojek Users

Type	Probabilitas
Positif	0,084033613
Negatif	0,504201681

The final probability is 0.084033613 for positive, and 0.504201681 through Naive Bayes calculations. It can be concluded that there are more negative tweets than positive tweets. And also the percentage of occurrences of negative words in table 4, the result is much larger than the percentage of occurrences of positive words.

3.1. Grab Data Analysis

The following is the crawled data for the keyword 'Grab' using the Python programming language.

Table 5. Percentage Of Grab User Sentiments

Tweet	Type Tweet		Type Tweet (%)	
	Positif	Negatif	Positif (%)	Negatif (%)
1	[]	[' slow']	0	20

Tweet	Type Tweet		Type Tweet (%)	
	Positif	Negatif	Positif (%)	Negatif (%)
2	[]	[slow]	0	7,142857143
3	[]	['disappointed', slow]	0	16,66666667
4	['fast']	['slow']	4,347826087	4,347826087
5	[]	[disappointed', slow]	0	14,28571429
6	['fast']	['slow']	4	4
7	[]	['congested']	0	7,692307692
8	[]	['disappointed', slow]	0	14,28571429
9	[fast"]	['slow]	4	4

Table 6. Results of Respondents' Sentiment of Grab

Type	Positif	Negatif
Probabilitas	0,3333333333	1
Mean	1,371980676	10,26900957

There are 9 tweets with each classification. There are 3 positive tweets with a probability of 0.333333333, and a negative tweet with a probability of 1. For the percentage for each type of tweet, it is 1.371980676% for positive, and 10.26900957% for negative. For classification related to word occurrence data, it is in the following table.

Table 7. Opinion Factors for Grab User Data

Type	Word Occurrence	Probabilitas	
		Positif	Negatif
slow	8	0	0,53
disappointed	3	0	0,2
fast	3	0,2	0
congested	1	0	0,07
Total	15	0,2	0,79

In table 7, there are occurrences of words with their respective numbers. The probability of word occurrence is the result of each word occurrence among the total number of words that appear in the tweet, which is 15. The probability result of a positive word occurrence is 0.2 and a negative one is 0.79. The following is the result of calculating the total probability using the Naive Bayes method.

Table 8. Results Of Calculation Of Total Probability Users Of Grab

Type	Probabilitas
Positif	0,066666667
Negatif	0,79

The final probability is 0.66666667 for positive, and 0.79 for negative through Naive Bayes calculations. It can be concluded that there are more negative tweets than positive tweets. And also the percentage of occurrences of negative words in table 8, the result is

much larger than the percentage of occurrences of positive words. The following is the crawled data for the keyword 'Commuter Line' using the Python programming language:

Table 9. Percentage of Commuter Line User Sentiment

Tweet	Type Tweet		Type Tweet (%)	
	Positif	Negatif	Positif (%)	Negatif (%)
1	[cheap]	[]	4,545454545	0
2	['fast]	[]	10	0
3	['good]	[]	20	0

Table 10. Sentiment Results of Commuter Line Respondents

Type	Positif	Negatif
Probabilitas	1	0
Mean	11,51515152	0

There are 3 tweets with each classification. There are 3 positive tweets with a probability of 1 and a negative tweet with a probability of 0. The percentage for each type of tweet is 11.5151515% for positive, and 0% for negative. It can be seen in the probability and percentage that the positive value generated is greater, which means that there are more positive tweets than negative tweets. Not even negative tweets were generated from the data we processed. For classification related to word occurrence data, it is in the following table.

Table 11. Opinion Factors for Commuter Line User Data

Type	Word Occurrence	Probabilitas	
		Positif	Negatif
Cheap	1	0,33	0
Fast	1	0,33	0
Good	1	0,33	0
Total	3	0,33	0

In table 11, there are occurrences of words with their respective numbers. The probability of word occurrence is the result of each word occurrence among the total number of words that appear in the tweet, which is 3. The probability result of a positive word occurrence is 0.33 and a negative one is 0. The following is a total probability calculation result using the Naive Bayes method.

Table 12. Result of Total Probability Calculation of Commuter Line Users

Type	Probabilitas
Positif	0,333333333
Negatif	0

The final probability is 0.333333333 for positive, and 0 for negative through Naive Bayes calculation. It can be concluded that there are more negative tweets than positive tweets. And also the percentage of occurrences of negative words in table 12 is much greater than the percentage of occurrences of positive words.

3.2. Transjakarta Data Analysis

The following is the crawled data for the keyword 'Transjakarta' using the Python programming language.

Table 13. Sentiment Percentage of Transjakarta Users

Tweet	Type Tweet		Type Tweet (%)	
	Positif	Negatif	Positif (%)	Negatif (%)
1	[]	['lazy']	0	4,545454545
2	['good']	[]	20	0
3	[]	['slow', 'slow']	0	12,5
4	[]	['disappointed']	0	5,555555556
5	['good']	[]	7,692307692	0
6	[]	['slow']	0	4,545454545

Table 14. Sentiment Results of Transjakarta Respondents

Type	Positif	Negatif
Probabilitas	0,333333333	0,666666667
Mean	4,615384615	4,524410774

There are 6 tweets with each classification. There are 2 positive tweets with a probability of 0.333333333, and a negative tweet with a probability of 0.666666667. The percentage for each type of tweet is 4.615384615% for positive, and 4.524410774%. It can be seen in the probability and percentage that the resulting negative value is greater, which means that there are more negative tweets than positive tweets. For classification related to word occurrence data, it is in the following table.

Table 15. Opinion Factors of Transjakarta User Data

Type	Word Occurrence	Probabilitas	
		Positif	Negatif
Lazy	1	0	0,14
Good	2	0,29	0
Slow	3	0	0,43
Disappointed	1	0	0,14
Total	7	0,07	0,18

In table 15, there are occurrences of words with their respective numbers. The probability of word occurrence is the result of each word occurrence among the total number of words that appear in the tweet, which is 6. The probability result of a positive word occurrence is 0.07 and a negative one is 0.18. The following is the result of calculating the total probability using the Naive Bayes method.

Table 16. Result of Total Probability Calculation of Transjakarta Users

Type	Probabilitas
Positif	0,023809524
Negatif	0,119047619

The Final probability is 0.023809524 for positive, and 0.119047619 through Naïve Bayes calculations. It can be concluded that there are more negative tweets than positive tweets. And also the percentage of occurrences of negative words in table 16, the result is much larger than the percentage of occurrences of positive words.

3.3. Implications of the Fourth Analysis Results

From the observations and analysis that the researchers did, the researchers wanted to prove what transportation services were considered good and which were considered bad by the Twitter user community. The value of 'good' is taken from the results of the percentage of occurrences of positive words that the researcher has analyzed in each of the points above and the value of 'bad' is taken from the results of the percentage of occurrences of negative words from data collection and calculations that have been carried out.

3.4. Criticism and Suggestions

From the results of observations and analyzes that have been carried out, it is hoped that it will be useful for each organization in monitoring service reviews submitted by the public on Twitter social media. And this research is also expected to be used as a consideration to improve the public transportation services of each organization in the future. The suggestion from the researcher is that it would be better for each public transportation service organization to improve the services provided on Twitter social media so that the reviews given by the public or customers can be monitored and the relationship to customers will be more interactive.

4. Conclusion

From the results of data collection carried out for 5 hours on each keyword and also carried out calculations using the Naïve Bayes method, it was resulted that the Commuter Line transportation service had the highest positive final probability, namely 0.333333333. Next is Gojek which has a positive ending probability of 0.084033613. While the positive final probability for Grab transportation services is 0.066666667. And finally, the final positive probability for Transjakarta transportation services is 0.023809524. There is also a negative final probability in the results of the analysis of this study. The first one was obtained by Grab, which was 0.79. Next is Gojek which has a negative final probability of 0.504201681. Then Transjakarta which has a negative final probability of 0.119047619. And finally, the Commuter Line which does not have negative tweets, which means it has a probability of 0. Based on the system built, the total positive final probability sentiment result is 0.507843137, and the negative final probability sentiment result is 1.4132493. The results show that the level of negative sentiment from public tweets is greater than the level of positive sentiment.

References

- [1] Kasmir, "Manajemen Perbankan", PT.Raja Grafindo, Jakarta, pp. 22, 2010. (references)
- [2] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis", *Procedia Computer Science*, pp. 26-32, 2013.
- [3] Nasukawa, and Yi, "Classification of Local News Using the Naïve Bayes Method with the N-Gram Feature. *ASIA Scientific Journal of Technology and Information (JITIKA)* Vol. 10, No. 1, Februari, 2003
- [4] Cvijikj, and Michahelles, "Combining Classification and Clustering for Tweet Sentiment Analysis", *Journal of Computational Linguistics and Chinese Language Processing*, December, 2011

-
- [5] Troussas et al., “Sentiment Analysis Capturing favorability using Natural Language Processing”, Proceedings of the National Seminar on Science & Technology Application (SNAST) Period III, 2013
 - [6] Olsen, D. L., Delen, D., Advanced Data Mining Techniques, Penerbit pringer, USA, 2008.
 - [7] Olson, David L, and Delen, D, “Advanced Data Mining Techniques”, Springer, United States of America, 2008.
 - [8] Bustami, “Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah Asuransi”, TECHSI, Jurnal Penelitian Teknik Informatika Vol. 3, pp. 127-146, 2013
 - [9] Han Jiawei, Michelline Kamber, and Jian Pei, “Data Mining Concepts and Techniques”, Elsevier, United States of America, pp 84-117, 2012