Sentiment Analysis of Twitter Social Media to Online Transportation in Indonesia Using Naïve Bayes Classifier

Dini Fakta Sari, Deborah Kurniawati, Edy Prayitno, Irfangi
STMIK AKAKOM Yogyakarta, Indonesia

ABSTRACT

Social media has made it easier for people to express their opinions on various matters, including online transportation services that are growing rapidly. This research was conducted to find out public opinion on the use of online transportation services. The method for analyzing the Naïve Bayes Classifier method for analyzing social media Twitter sentiment towards online transportation in Indonesia. The study was conducted by processing 1009 data, which consisted of 900 training data and 109 test data. The test results, with an accuracy of 84%, showed 11% positive values, 14% negative values, and the remaining 75% were neutral values. In fact, it does not affect the user's decision to utilize online transportation services.

Keywords: naïve bayes classifier, online transportation, sentiment analysis, social media

INTRODUCTION

The rapid development of information technology has brought many changes in all fields in Indonesia, including the transportation industry, namely by developing various kinds of online transportation services. Online transportation is one of the latest service innovations in m-commerce. Online transportation services or travel sharing are individual transportation services where customers can order a ride (car, motorcycle, etc.) through a mobile application and the driver can respond to orders through the application (Wallsten, S., 2015). Online transportation is the answer to people's needs for transportation that is easy to get, convenient, fast, and cheap. This makes online transportation very much needed by many people to fulfill their various transportation needs. Online transportation provides several benefits such as the driver and the customer can find out the location of each other accurately; the customers can see driver and vehicle information; and the customers can easily find transportation to go to another place (time efficiency) (Farin, N. J., Rimon, M. N. A. A., Momen, S., Uddin, M. S., & Mansoor, N. (2016)). These benefits make ride-sharing gain popularity among urban people easily (Shilvia L. Br. Silalahi, Putu W. Handayani, & Qorib Munajat, 2017). There are several online transportation services that are popular in Indonesia, namely GO-JEK, Grab, Uber, Bajaj App, Transjek, Wheel Line, Bangjek, Ojek Syar'I, and Blue-Jek (Okezone.com, 2015).

The custom of today's society is to express their opinions through social media. All things can be discussed in social media, one of which is online transportation. One social media that is widely used by people to express their opinions is Twitter. With Twitter the public can show positive responses or emotions about online transportation. By giving a tweet via Twitter, the public can also reveal their income about online transportation from a negative perspective. But people can also not show positive or negative expressions in their tweets, but neutral.

Online transportation services have unique characteristics among other m-commerce services in terms of the nature of their services. It has a different service process, involving inseparable physical services (eg drivers, vehicles, etc.) and the tendency of users to use services repeatedly. Because it's

important to learn about online transportation services (Shilvia L. Br. Silalahi, Putu W. Handayani, & Qorib Munajat, 2017).

Meanwhile, consistent growth from users and user-generated content on many websites, social networks and online consumer platforms such as Twitter, Amazon, and Yelp has increased the amount of information available on the internet (Oscar Araque, Ganggao Zhu, & Carlos A. Iglesias, 2019).

Sentiment analysis is centered around the classification of sentiments, opinions or expressions expressed in human-generated texts. For this purpose, text can be labeled into several categories, being the most common positive and negative (Oscar Araque, Ganggao Zhu, & Carlos A. Iglesias, 2019). Sentiment analysis uses the natural language processing (NLP), text analysis and computational techniques to automate the extraction or classification of sentiment from sentiment reviews (Basant A., Namita, M., Pooja, B., & Garg, S., 2015). Analysis of these sentiments and opinions has spread across many fields such as consumer information, marketing, books, application, websites, and social (Doaa Mohey El-Din Mohamed Hussein, 2016).

One method for analyzing sentiment is Naive Bayes Classifier (NBC). The Naïve Bayesian machine learning approach is one of the most popular methods used in predictive analytics. NBC converges quicker than discriminative models like logistic regression, as it needs lesser training data comparatively. (Sandhya Maitra, Sushila Madan, Rekha Kandwal, & Prerna Mahajan, 2018). NBC is a simple probabilistic based prediction technique based on the application of the Bayes theorem with the assumption of strong or naive independence. NBC is a simple method but has high accuracy and performance in text classification.

This research was conducted to find out trends or sentiments of the Twitter user community in expressing their opinions about online transportation through their tweets. Do they give positive, negative or neutral opinions about existing online transportation services?

METHOD

This research was conducted using 2 data that must be present for input needs in classification 2, namely, tweets that have known sentiment values and tweets whose sentiment value is unknown. The tweet data containing the word "gojek motor" will be saved first into the collection on the JSON (JavaScript Object Notation) file. The word "gojek motor" is an Indonesian word meaning "motorbike taxi". The data is obtained from 2017 to 2018 which will then be used for preprocessing purposes. Block diagram of the system can be seen in Fig.1 below.

From Fig. 1 above it can be explained that, input is done by retrieving data tweets based on tweets that contain the word "gojek motor". The data will be used for the next stage. In the process phase , several processes are carried out as follows, namely Pre-Processing which includes Case-Folding, Tokenizing, and Filtering. In the Case-Folding phase, all capital letters of the tweet will be changed to lowercase, and deletion of all punctuation marks such as commas, periods, question marks and so on will be done. In the Tokenizing phase, the word spacing will be separated in the tweet, then the separated words will be used for the next stage. And at the filtering phase will be deleted the words that are not used.

Data that has been pre-processed, can then be used as test data and training data. Training data will be classified into classes that have been determined which are negative, positive and neutral. So at this phase it can be regarded as weighting each word. Meanwhile, at the stage of classification of the test data, the test data will be carried out on a probability calculation process based on training data. From the classification process, the probability data of the test data for the negative, positive and neutral classes will be obtained. The highest probability value will be determined as the sentiment value of the test data.

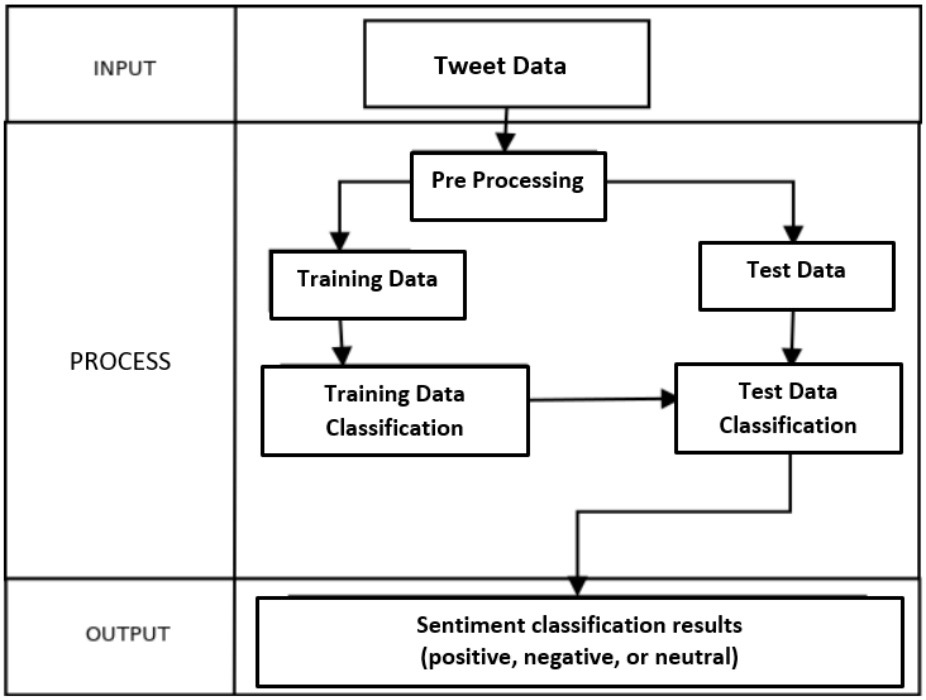

Fig. 1. Block diagram of the system

The system developed has several functions which can be accessed by users, namely retrieving tweet data; pre-processing the tweet data; determine the training data class; NBC testing; and display the test results.

After the data is taken, tweet data will be saved in JSON format, and the pre-processing will be done to determine the weight or class of tweets. The weighted data is used for the next process, namely training using the NBC. After passing this process, the data will be divided into training data and test data. Training data is determined by the user's sentiment class. After that the data is ready to be used in the next process. When testing test data will be calculated based on previous training data. And as a result, the highest probability value will be obtained as a conclusion of the sentiment value that will be displayed.

The probability of each word is calculated using the following formula

$$P(Wk|Vj) = \frac{nk+1}{n+|vocabulary|}$$

Where *nk* is the number of times the occurrence of each word appears, *n* is the number of frequencies of occurrence of words from each category, |vocabulary| is the number of all words from all categories.

RESULT AND DISCUSSION

The data used in this study was obtained from twitterscraper which will be saved into JSON files. In the JSON file the text variable value is taken to be used in the next process. The contents of the JSON file are FullNames that contain the user name used; html which contains html scraping tags; tweet id; likes which shows the number of likes on the tweet; replies which means how many times the tweet is returned; timestamp, contains the time when the tweet was made; text, the contents of the tweet; the url of the tweet link; and the user who tweeted.

The existing tweet data is weighted into 3 categories, namely positive, negative and neutral done manually, and then stored in the form of a txt file. Examples of pre-processing training data can be seen in Table 1 below.

Table 1. Examples of training data

Data	Tweet	Pre-processing
1	Driver gojek sekarang motornya jelek dan naik motornya ngebut #ojol @gojek_indonesia	Driver, gojek, sekarang, motornya, jelek, dan, naik, motornya, ngebut, ojol, gojek_indonesia
2	Terimakasih @gojek_indonesia dapat promo, voucher gratis dan dapat driver sopan lagi.	Terimakasih, gojek_indonesia, dapat, promo, voucher, gratis, dan, dapat, driver, sopan, lagi.

Next, the sample data that has been preprocessed is calculated by the frequency of its appearance, as shown in Table 1 and Table 2.

Table 2. Frequency of occurrence of the word data 1

No	Word	Appearance	Sentiment
1	Driver	1	Neutral
2	Gojek	1	Neutral
3	Sekarang	1	Neutral
4	Motornya	2	Neutral
5	Jelek	1	Negative
6	Dan	1	Neutral
7	Naik	1	Neutral
8	Ngebut	1	Negative
9	Ojol	1	Neutral
10	Gojek_indonesia	1	Neutral

Table 3. Frequency of occurrence of the word data 2

No	Word	Appearance	Sentiment
1	Terimakasih	1	Positive
2	Gojek	1	Neutral
3	Indonesia	1	Neutral
4	Dapat	2	Positive
5	Promo	1	Positive
6	Voucher	1	Positive
7	Dan	1	Neutral
8	Driver	1	Negative
9	Sopan	1	Positive
10	Lagi	1	Neutral

From the two tables above, it can be seen that the frequency of negative occurrences is 2 times, the frequency of positive appearances is 5 times, and the frequency of neutral occurrence is 13 times.

Table 4. Frequency of occurrence of data for each category

Category	Example		Total	Arg Max
	Data 1	Data 2		
Positive	0	1	1	½
Negative	1	0	1	½
Neutral	1	1	2	2/2

Table 5. Calculation of probability of each word

Word	**Prob. Positive**	**Prob. Negative**	**Prob. Neutral**
Driver	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{2+1}{13++20}=\frac{3}{33}$
Gojek	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{1+1}{13++20}=\frac{2}{33}$
Sekarang	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{1+1}{13++20}=\frac{2}{33}$
Motornya	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{2+1}{13++20}=\frac{3}{33}$
Jelek	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$
Dan	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{2+1}{13++20}=\frac{3}{33}$
Naik	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{1+1}{13++20}=\frac{2}{33}$
Ngebut	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$
Ojol	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{1+1}{13++20}=\frac{2}{33}$
Terimakasih	$\frac{0+1}{5++20}=\frac{2}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$
Gojek_indonesia	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{2+1}{13++20}=\frac{3}{33}$
Dapat	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{1+1}{13++20}=\frac{2}{33}$
Promo	$\frac{0+1}{5++20}=\frac{2}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$
Voucher	$\frac{0+1}{5++20}=\frac{2}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$
Gratis	$\frac{0+1}{5++20}=\frac{2}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$
Sopan	$\frac{0+1}{5++20}=\frac{2}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$
Lagi	$\frac{0+1}{5++20}=\frac{1}{25}$	$\frac{0+1}{2++20}=\frac{1}{22}$	$\frac{0+1}{13++20}=\frac{1}{33}$

The probability above is then used to calculate the sentiment of the test data. Then the calculation is based on the training data that has been calculated before.

Table 6. Example of testing data

Tweet	Preprocessing
Pagi ini dapat voucer gratis trimakasih @gojek_indonesia	Pagi, ini, dapat, voucer, gratis, trimakasih, gojek_indonesia

Vmap(positive) = P(“positive”)P(”dapat”| positive)P(“voucer” | positive) P(“gratis” | positive) P(“trimakasih” | positive) P(“gojek_indoneisa” | positive) =

$$\frac{1}{2} \times \frac{1}{25} \times \frac{2}{25} \times \frac{2}{25} \times \frac{2}{25} \times \frac{1}{25} = \mathbf{0.000000409}$$

Vmap(negative) = P(“negative”)P(”dapat”| negative)P(“voucer” | negative) P(“gratis” | negative) P(“trimakasih” | negative) P(“gojek_indoneisa” | negative) =

$$\frac{1}{2} \times \frac{1}{25} \times \frac{1}{25} \times \frac{1}{25} \times \frac{1}{25} \times \frac{1}{25} = 0.000000051$$

Vmap(neutral) = P(“neutral”)P(”dapat”| neutral)P(“voucer” | neutral) P(“gratis” | neutral) P(“trimakasih” | neutral) P(“gojek_indoneisa” | neutral) =

$$1 \times \frac{2}{25} \times \frac{1}{25} \times \frac{1}{25} \times \frac{1}{25} \times \frac{2}{25} = 0.000000408$$

The sentiment value from the results of the calculation above is the one that has the largest Vmap value, which is "Positive".

In this study using 1010 data which was then taken 900 data for training purposes and 110 data for testing purposes. Then 10% has a positive value of 14% has a negative value and 76% has a neutral value.

From the accuracy test conducted, it can be seen that there are 18 tweets that have sentiment values that are not as desired, then accuracy can be calculated as follows:

$$\text{Accuracy} = \frac{n - \ total\ error}{n} x\ 100\% = \frac{110 - 18}{110} x\ 100\% = 84\%$$

So it can be said that the accuracy of the system created is 84%

CONCLUSION

This study uses the Naïve Bayes Classifier method to analyze the sentiments of community tweets regarding online transportation, and group them into 3 categories namely positive, negative, and neutral. From the analysis of 1010 data then 900 data were taken for training purposes and 110 data for testing purposes, 10% had a positive value of 14%, had a negative value and 76% had a neutral value, with system accuracy of 84%

ACKNOWLEDGMENT

Through this paper, we would like to express our gratitude to all parties who have supported the implementation of this research, especially the lecturer friends who have provided additional information about the material and references, and students at STMIK AKAKOM Yogyakarta who assisted in the development of the system. Thank you for all the kindness.

REFERENCES

Araque, O., Zhu, G., and Iglesias, C.A., 2019, A semantic similarity-based perspective of affect lexicons for sentiment analysis, Knowledge-Based Systems 165, 346–359.

Basant, A., Namita, M., Pooja, B., and Garg, S., 2015, Sentiment Analysis Using Common-Sense and Context Information. Hindawi Publishing Corporation Computational Intelligence and Neuroscience.

Farin, N. J., Rimon, M. N. A. A., Momen, S., Uddin, M. S., and Mansoor, N. (2016). A framework for dynamic vehicle pooling and ridesharing system. In Computational Intelligence (IWCI), International Workshop on (pp. 204-208). IEEE.

Hussein, D.M.E.M., 2016, A survey on sentiment analysis challenges, Journal of King Saud University – Engineering Sciences *30*, 330–338

Maitra, S., Madan, S., Kandwal, R., and Mahajan, P., 2018, Mining authentic student feedback for faculty using Naïve Bayes classifier, Procedia Computer Science 132, 1171–1183.

Okezone.com (2015). 10 Jasa Transportasi Online di Indonesia, dari Gojek hingga Uber. September 23, 2015. http://economy.okezone.com/read/2015/09/23/320/1219859/10-jasa-transportasi-online-di-indonesia-dari-go-jek-hingga-uber

Shilvia L. Br. Silalahi, Handayani, P.W., and Munajat, Q., 2017, Service Quality Analysis for Online Transportation Services: Case Study of GO-JEK, Procedia Computer Science 124, 487–495.

Wallsten, S. (2015). The competitive effects of the sharing economy: how is Uber changing taxis. Technology Policy Institute, 22.