Classification of community opinion on the use of the Transjakarta bus based on twitter social network using naïve bayes method

B. D Meilani¹, R K Hapsari², I F Novian¹

¹Department of Information System, Faculty of Electrical and Information Technology, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
²Department of Informatics, Faculty of Electrical and Information Technology, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia

budanis@itats.ac.id

Abstract. At this time, social networking is becoming a very popular communication tool among Indonesians ranging from children to adults. One of them is Twitter, which has quite a lot of users. This social network is used by users to write their opinions on a matter. One of the topics discussed in the Twitter social network is the Trans-Jakarta bus. In this study, the opinions of Trans-Jakarta bus users listed on Twitter are only limited to positive and negative opinions. An application designed to predict positive or negative opinions using the Naïve Bayes algorithm. The Naïve Bayes algorithm was chosen because it only requires a small amount of training data to estimate averages and variable variations. The developed application can make it easier for decision-makers towards Trans-Jakarta to find out public opinion on Trans-Jakarta bus services. The resulting system accuracy is 73% because the amount of training data used is only 62.5% of the total data of 50 data. While the test data that has been used are only 30 data. The proportions of the training and testing data affect the level of accuracy, so if you want a higher level of accuracy you can reproduce the training data used.

1. Introduction
Currently, the Intelligent Transport System (ITS) is aggressively developing urban transportation that is comfortable and easily accessible to the public. With this public transportation, all levels of society can take advantage. So that it can unravel congestion which has been a problem in big cities like Jakarta. "Based on BPTJ data, currently the average speed of vehicles in busy times is 10 km/hour, decreasing every year and causing economic losses. The impact of traffic congestion in Jakarta alone reaches the US $ 1 billion in 2010 and is expected to increase to the US $ 6.5 billion by 2020 if this condition is not corrected" [1].

To overcome traffic congestion in Jakarta, the transfer of the use of private transportation to use public transportation. One of them is the public transportation Trans-Jakarta bus which is the transportation choice of the people of Jakarta. "Our target for 2017 is 185 million people, meaning that it will increase by 50%, the increase in the number of passengers will increase every day" [2]. In 2015 there were 207,000 daily Trans-Jakarta bus passengers and the number increased by 50% in 2016. With the increase in Trans-Jakarta users, many users gave their opinions. Public opinion on Trans-Jakarta transportation is very much, wherein one month there are about ± 100 opinions that enter. The opinion
is in the negative and positive classification. With the existence of public opinion towards Trans-Jakarta, the Trans-Jakarta party can find out the level of community satisfaction with trans-Jakarta transportation. To get an opinion on Trans-Jakarta one way is to use social media networks, namely Twitter. At the official Twitter development conference, management delivered statistics on the site and Twitter users, the statistics stated that in April 2010, “Twitter has 106 million accounts and as many as 180 million unique or different users each month. The number of Twitter users has 106 million accounts and an increase of 300,000 users every day” [3].

Research in the field of opinion mining began to bloom in 2002. Turney in 2002 conducted research on opinion mining by using data in the form of consumer review data on a product. By implementing the Semantic Orientation method using Pointwise Mutual Information (SO-PMI). The best results were achieved, with an accuracy of 84% of the motor vehicle review data and an accuracy of 66% for the film review data [4]. While research on the classification of sentiment towards the review of the movie has been done by Dhande and Patnaik using the Naive Bayes algorithm, Neural Network, and Naive Bayes classifier Neural. From the results of the final research that was tested using the three algorithms mentioned that Naive Bayes produces higher accuracy than Neural Network [5].

Naive Bayes is one of the methods in the classification of documents. Naive Bayes is a classification with probability and statistical methods proposed by the British scientist Thomas Bayes. At this time the classification of text using the Naïve Bayes algorithm has experienced many improvements such as reducing the dimensions of word features to improve algorithm performance. In addition to the small training data, rare word features appear randomly. The research found that the precision results from 80% increased to 85%, the recall ratio from 81% increased to 83%, and the f-measure value in the classification from 81% increased to 84% [6].

Many successful studies using the Naive Bayes method include Data Mining Applications to produce Student Graduation Patterns with the Naive Bayes Method [7] and Classification Identification of Factors that Cause Inaccuracy in Students Graduation with the Naive Bayes Classifier Method [8].

2. Method
Data mining is an interdisciplinary field that is part of the process of finding greater knowledge. In data mining includes pre-processing tasks such as extraction, cleaning, mixing, data reduction and construction features as well as post-processing steps such as pattern and model interpretation, hypothesis confirmation and generation [9].

2.1. Texts Mining
Text mining is an activity to retrieve data where the source of data is obtained from Word documents, PDFs, quoted text, which is obtained by extracting information, tracking topics, summarizing, categorizing, grouping, and linking concepts [10]. In-text mining, the process of taking information from textual data that has high quality and can find out the problems in the text of a particular topic. Text mining in sentiment analysis can identify emotionally about a statement [11]. Sentiment analysis often also known as opinion mining is a computational study of opinions, sentiments, attitudes, and emotions that are presented in a text [12].

According to Feldan and Sanger text mining can be considered a new research subject. Text mining can provide solutions to problems such as processing, organizing or grouping, and analyzing large amounts of unstructured text [10].

Text mining can also be assumed as an extension of the data mining or knowledge-discovery database (KDD), which is used to find interesting patterns of large data. But text mining has a more specific field than data mining because data is in the form of unstructured text. Besides text mining also refers to the discipline of computer science that is used to deal with natural language such as exploiting techniques and methodologies from the field of information search. Text mining relies on information retrieval, natural language processing, and information extraction [13].
2.2. Texts Preprocessing 

Pre-processing data is the process of cleaning and preparing texts for classification [14]. Online text usually contains a lot of noise. So it is necessary to do the text preprocessing stage. This stage is the stage where the data is prepared to be data that is ready to be analyzed, which aims to obtain satisfactory results. This process also aims to eliminate problems that can affect the results of the data mining process.

Preprocessing is an important part of this research system because the results obtained depend on whether or not we do pre-processing data. Preprocessing data will be very important to use on social networking data because most of the sentences used are not formal, unstructured, and have a large noise. To get good results, the raw data will be preprocessed in various stages. The preprocessing stage is shown in Figure 1, which consists of four stages, namely [15]:

1. Case Folding
   - The process of the folding case that converts the entire text in the document into lowercase.

2. Cleansing
   - The cleansing process is removing characters that are considered invalid, invalid characters include words preceded by characters ('@'), hashtags ('#'), links ('https') and symbol characters such as (~!@#$%^&*()-_=+{|}[]\|;:'"",>.?)/.

3. Tokenizing
   - The tokenizing process is cutting a document or sentence into parts of words, the sentence to be cut is a sentence separated by a space character, the process aims to facilitate the weighting of each word that appears.

4. Filtering
   - The filtering process is taking important useful words and conjunctions that do not count in the number of word occurrences.

**Figure 1. Data preprocessing flow**

2.3. Naïve Bayes

Naïve Bayes is a method proposed by a British scientist named Thomas Bayes. This technique is a probability and statistic that can predict future opportunities based on previous events, with a strong assumption of independence from each condition. Naïve Bayes for each class of decisions calculates the probability of condition that the class of decisions is correct. Given the vector of object information, this algorithm assumes that the object's attributes are independent [16].

Naïve Bayes is used in the classification process of documents to be classified into positive and negative classes. In this research, the dataset used is data obtained from Twitter after being processed first. The dataset is divided into two data, namely training data and testing data. Training data is data that the classes have already known, while testing data is data whose classes are not yet known.
The process flow in the main system is shown in Figure 2a, while the details of the Naïve Bayes classification sub-process are shown in Figure 2b. The naïve Bayes algorithm process begins by calculating the prior value which is the probability of class appearance compared to the total number of classes. The calculation of the prior value is written in formula 1.

\[
P(c_i) = \frac{f(d(c_i))}{|D|}
\]

where,
- \( c_i \) = document category
- \( f(d(c_i)) \) = number of documents that have \( c_i \) category
- \( D \) = the total number of training documents

The prior calculation is done for as many classes as there are. Then count the number of unique terms in the training data from all classes and count the appearance of each term in the document, by counting the number of occurrences of the word in each category \((c_i)\). The next calculated value is the number of words in each class, the words that are counted are not unique words but all the words in the category. After getting these values, the next step is to calculate the probability of each word against each existing category. The calculation formula is written in formula 2.

\[
P(w_k|c_i) = \frac{(nK+1)}{(n+(vocab))}
\]

where,
- \( w_k \) = words that come out in the \( c_i \) category
- \( nK \) = the number of occurrences of words in the \( c_i \) category
- \( n \) = the total number of words in the \( c_i \) category
- \( vocab \) = total number of words

3. Result
In testing this research used 30 test data. As well as a comparison between the results of the classification system and the results of the classification which is done manually. The test data is taken randomly from the dataset. The test results of the 30 opinion data are shown in Table 2.
Based on the results of testing of the testing data, it is used as a reference to calculate the accuracy of the system. Accuracy is calculated by formula 3.

\[
\text{accuracy} = \frac{\text{the amount of correct data}}{\text{amount of testing data}} \times 100\% \tag{3}
\]

The level of accuracy is obtained from:

\[
\text{accuracy} = \frac{22}{30} \times 100\% = 73\%
\]

So that the resulting accuracy is 73%.

The resulting system accuracy is 73% because the amount of training data that has been used is only 62.5% of the total data, namely 50 data. While the test data that has been used are only 30 data. The proportions of the training and testing data affect the level of accuracy, so if you want a higher level of accuracy you can reproduce the training data used.
4. Conclusions

Based on the development of the Trans-Jakarta user opinion classification application on the Twitter social network using the Naïve Bayes algorithm, it can be concluded that: The classification application has been able to classify 30 test data with an accuracy of 73%, and this application has made it easier for Trans-Jakarta directors to find out public opinion about Trans-Jakarta buses so as to improve the quality of Trans-Jakarta buses.

5. References

[1] P. Noni, 2017, “ITS Indonesia : penerapan teknologi untuk integrasi transportasi sudah mendesak,”. [Online]. Available: Tribunnews.com.
[2] Joseph, 2017. “Pengguna Transjakarta di Tahun 2017,” Kompas.com, Jakarta.
[3] Yarrow, M. Radke, J. A. Clausen and P. R. Robbins, 2010, “The Social Meaning of Mental Illness,” Journal of Social Issues, vol. 11, no. 4, pp. 33-48.
[4] P. D. Turney, 2002, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” in the Association for Computational Linguistics 40 Anniversary Meeting, New Brunswick, N.J.
[5] L. L. &. P. P. G. K. Dhande, 2014 “Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier,” International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 3, no. 4, pp. 313-320.
[6] L. Yuan, 2010, “An Improved Naive Bayes Text in Chinese Information Processing,” Henan Polytechnic University, Jiaozuo, China.
[7] B. D. Meilani and N. Susanti, 2015, “Aplikasi Data Mining untuk Menghasilkan Pola Kelulusan Siswa dengan Metode Naive Bayes,” Jurnal Ilmiah NERO, vol. 1, no. 3.
[8] B. D. Meilani, S. Wahyudiana, A. Y. P. Putri and A. Pakarbudi, 2019, “Klasifikasi Identifikasi Faktor Penyebab Ketidaktepatan Masa Lulus Mahasiswa dengan Metode Naive Bayes Classifier,” in Seminar Nasional Sains dan Teknologi Terapan, Surabaya.
[9] M. J. &. M. J. W. Zaki, 2014, Data Mining and Analysis : Fundamental Concepts And Algorithms, New York: Cambridge University Press.
[10] R. &. S. J. Feldman, 2007, The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data, New York.: Cambridge University Press.
[11] A. Z. Khan, M. Atique and V. Thakare, 2015, “Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis,” International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), pp. 89-91.
[12] L. B, “Sentiment Analysis and Opinion Mining, 2012,” Synthesis Lectures On Human Language Technologies, vol. 5, no. 1.
[13] A. Hotho, A. Numberger and G. PaaB, 2005, “ A Brief Survey of Text Mining,” Computational Linguistics and Language Technology.
[14] E. Haddi, X. Liu and L. Shi, “The Role of Text Pre-processing in Sentiment Analysis, 2013,” First International Conference on Information Technology and Quantitative Management, pp. 26-32.
[15] S. Aji, 2017, “Sistem Klasifikasi Opini Pengguna Maskapai Penerbangan Indonesia Pada Jejaring Sosial Twitter Menggunakan Metode K-Nearest Neighbor,” Malang.
[16] O. &. Delen, Advanced Data Mining Techniques, 2008, USA : Springer-Verlag Berlin Heidelberg.