# SENTIMENT ANALYSIS OF UINSU STUDENTS' COMFORT TOWARDS TRANS METRO DELI SERVICES AT TAMAN BUDAYA BUS STOP USING THE NAIVE BAYES METHOD

[1*]NURHIDAYATI, [2]DEA SYAHFIRA HASIBUAN, [3]LAILAN SOFINAH HARAHAP

[1,2]Department of Computer Sciences, Faculty of Sciences and Technology, Universitas Islam Negeri Sumatera Utara

Jl. William Iskandar Pasar V, Medan Estate, Kec. Percut Sei Tuan, Kab. Deli Serdang, Sumatera Utara, 20371

[3]Department of Information Technology, Universitas Muhammadiyah Sumatera Utara

Jl. Kapten Mukhtar Basri No. 3, Medan, Sumatera Utara, 20238

e-mail: [1]hnur87531@gmail.com, [2]deasyahfira16@gmail.com, [3]lailansofinahharahap@umsu.ac.id

[*]Corresponding author

ABSTRACT

*The large number of bus passengers at the Taman Budaya bus stop is one of the public transportation problems. Finding that the Metro Deli Bus Organizer is still operating. They are considered capable of meeting the requirements for choosing a mode of transportation. The purpose of this study is to determine the passenger transportation factors on the Trans Metro Deli Bus. The Trans Metro Deli Bus Passenger Transportation Factor is the purpose of this study. Data Collection Techniques Using Questionnaires Student comfort factors when using the Trans Metro Deli bus service. This study's methodology starts with problem identification and moves on to problem-solving techniques and assessment procedures. Respondents were given a questionnaire to fill out to collect data. The author of this study used Google Forms. The author of this study solved the problem using the Naïve Bayes algorithm. The Naïve Bayes algorithm model produces results with an accuracy of up to 71.43%, which is quite good. The accuracy results of 71.43% and approaching 100% show how accurate the sentiment analysis is using the Naïve Bayes classification. The accuracy results of 71.43% and approaching 100% show how accurate the sentiment analysis is using the Naïve Bayes classification. 'The bus took a long time to arrive' and 'didn't get a seat' were the most common negative reviews, indicating that some students felt uncomfortable. The Naïve Bayes results of the study showed that people who reviewed the Trans Metro Deli Bus expressed more positive opinions, with the highest score of 71.43%.*

**Keywords**: *Trans Metro Deli Bus, Naïve bayes, Sentiment analysis, Classification, Transportation*

## 1. INTRODUCTION

Along with the increasingly dense urban life, of course the demands of the Medan city community have a high interest in consumer decisions to use this bus transportation when carrying out daily activities. In 2023, the total number of Trans Metro Deli bus consumers reached 3,047,358 users. In March, the number of bus users reached 321,676 users, which is the highest number of passengers in 2023, and in July, bus passengers only reached 198,368 users, the lowest number of users in 2023. Based on the data, the number of bus users was lowest in July 2023, at only 198,368. Thus, it can be seen that in 2023, bus transportation consumers experienced instability in public interest in using Trans Metro Deli bus transportation. Consumer decisions in choosing to use a product or service have a significant role for the company because they can increase the profits obtained by the company in selling the products or services offered. Concerning consumer decisions, people use decision factors to choose which public transportation to use as a means of public transportation that supports daily activities in the city. Many factors can influence consumer decisions, such as service quality [1]. Price and quality of service are two major factors that consumers consider when making decisions.

Even today, public transportation still needs to be improved, which leads to a decrease in quality in terms of load frequency, speed, income, and comfort. This type of public transportation results from a system run by transportation

services that use deposit money, so it seems careless and does not prioritize passenger satisfaction. Therefore, Medan residents prefer private transportation for fast and practical travel. Compared to the public transportation system, which has not changed, the use of private transportation by the public will increase every year. As a result of the daily mobility of the community, it is predicted that the city of Medan will experience congestion (locked roads due to density at specific points) on several main roads in 2022 [2].

In previous studies, research was conducted to analyze user sentiment towards the performance of the public transportation system in Jakarta using the Naïve Bayes algorithm. Various studies have shown that sentiment analysis is effective in understanding public opinion regarding transportation services, with the use of the Naïve Bayes method proven to handle unbalanced data and achieve high accuracy. The results of this study indicate that the Naïve Bayes algorithm can be used to classify netizen comments on public transportation in the Jakarta area with a relatively high level of accuracy [3]. Based on previous research on sentiment analysis, which includes using Naïve Bayes to classify sentiment analysis from Indonesian online retailers, the Naïve Bayes approach can achieve a sentiment classification accuracy value above 90%. The 95% accuracy rate was achieved in other studies using Naïve Bayes, including the negation analysis feature. Sentiment analysis of telecommunications products and aviation services was also conducted using the Naïve Bayes method. With an accuracy rate of 89.50%, the study that highlighted the feature selection component was able to classify sentiment [4].

Over the past few years, many ideas regarding machine learning approaches to sentiment analysis problems have emerged, focusing on comparing three approaches. These techniques include Artificial Neural Networks, Decision Trees, and Naïve Bayes. The overall findings of this study suggest that the Naïve Bayes classifier is an ideal choice for domain training. Lexicon-based sentiment analysis techniques. This approach can produce low recall but high-precision [5]. Through rigorous testing and implementation, this study aims to demonstrate how this approach can improve the quality and comfort of student learning.

## 2. RESEARCH METHODOLOGY

### 2.1 Research Flowchart

We start with problem identification, problem-solving techniques, and evaluation processes; this study uses a methodical approach. Respondents were given a questionnaire to fill out to collect data. The author of this study used Google Forms. The author of this study solved the problem using the Naïve Bayes algorithm. The entire workflow is depicted in the figure above. Google Colab tools are used in data processing techniques to visualize the data and assess the classification after the data has been explored.

1. Crawling data
   Data retrieval by a crawler can unite data automatically. The required crawler can be adjusted to the characteristics of each platform; data crawling is done manually by researchers through Google Forms, and then the collected data is saved in Microsoft Excel format [6].
2. Processing data
   In this stage, researchers carry out processes such as data cleansing, the function of which is to remove data containing punctuation, emoticons, and numbering, then case folding to change words that start with capital letters to lower case letters, then tokenizing aims to separate each word in a sentence, then stemming to change abbreviated words into basic words. In this stage, researchers carry out processes such as:
   a. Cleansing data: its function is to remove data containing punctuation, emoticons, and numbering; case folding to change words that start with a capital letter to lowercase letters.
   b. Tokenizing aims to separate each word in a sentence. [7].
   c. Stemming to change words into basic words [8].

   Finally, the purpose of Filter Token (By Length) is to remove words consisting of 4 letters [9].
3. Labelling
   At the labeling stage, the researcher carries out the labeling process assisted by an expert in the field by holding discussions between both parties and identifying words containing positive or negative sentiments [10].
4. Splitting Data
   In the splitting data step, the data is divided into two parts: training data and test data. This stage is done by dividing the dataset into 80% training data and 20% of test data [11].
5. Implementation of Algorithm
   At this stage, the researcher implements the algorithm by processing data applied to a tool. Therefore, the researcher uses the Google Colab tool. The implementation steps are first to determine the machine learning algorithm that will be used, such as the naive Bayes algorithm, and then train the model using the training dataset. [12].
6. Conclusion
   At this stage, we can find out the words often used by netizens for this Public Transportation issue and the number of positive and negative class data obtained.

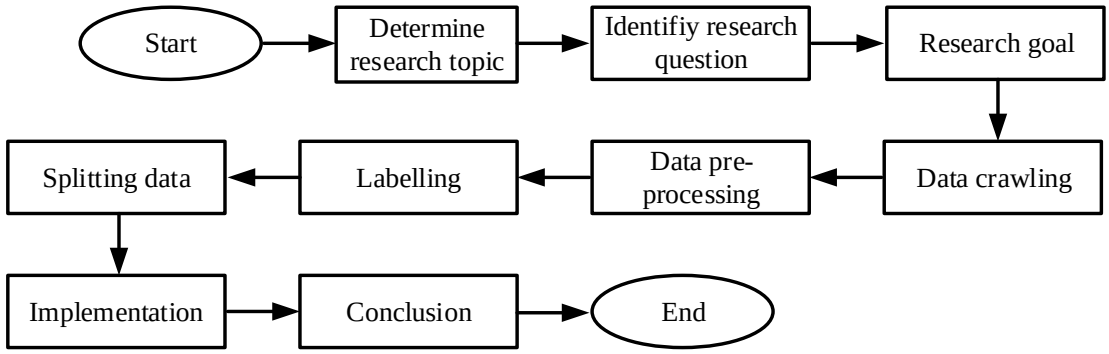


*Figure 1. Research Flowchart*

### 2.2 Implementation of Algorithm

One of the classification techniques that uses probability calculations is Naïve Bayes. Thomas Bayes originally proposed Bayes' Theorem, the basic idea behind the Naïve Bayes classifier. The following is a direct statement of the probability values used (Pop) [5]:

$$P\left(\frac{A}{B}\right) = \frac{P(A)P\frac{B}{A}}{P(A)} \quad (1)$$

where :
a: unique class
p(A/B): the possibility of hypothesis r based on condition s
p(A): r is the probability of the hypothesis
p(B/A): probability s according to the hypothesis r
p(B): probability s

According to Naïve Bayes, a document consists of its constituent words. The order in which words appear in a document is another thing that Naïve Bayes ignores. Table 1 shows the NBC algorithm for data classification.

## 3. RESULTS AND DISCUSSIONS

### 3.1 Crawling Data Results

In this study, researchers utilized the data crawling method and applied the Naïve Bayes model to classify data in sentiment analysis. Data was collected using a Google form to ensure its quality and accuracy. This analysis focused on three keywords: glass breaker, saving costs, and comfort, which were the primary focus for assessing student sentiment or responses to the Trans Metro Deli Service. The use of Indonesian in data collection allows the analysis to provide a more precise interpretation of the responses of Trans Metro Deli Service users, as presented in Table 2.

*Table 1. Classification data*

| | |
|---|---|
| Training data | (25, 76) |
| Test data | (7,76) |

*Table 2. Result of Crawling data*

| | | | | | |
|---|---|---|---|---|---|
| deasyahfira15@gmail.com | Ya | Ya | Ya | Pemecah kaca | Ya |
| yunilaras2003@gmail.com | Ya | Ya | Ya | Ya | Ya |
| diajengwulan2004@gmail.com | Ya | Tidak | Ya | Ya | Ya |
| maulizakhabira@gmail.com | Ya | Tidak | Ya | Terkadang saya dapat kursi, terkadang berdiri hingga halte terakhir | Ya |
| ihsantiara2127@gmail.com | Ya | Ya | Ya | Ya | Ya |
| serli22092003@gmail.com | Ya | Ya | Ya | Ya | Ya |
| juwitasar100104@gmail.com | Ya | Ya | Ya | pemecah kaca | Ya |
| mulfa7903@gmail.com | Ya | Ya | Ya | Ya | Ya |
| juwita10012004@gmail.com | Ya | Ya | Ya | Ya | Ya |
| manusiaorang98212@gmail.com | Ya | Ya | Ya | Ya | Ya |
| hastifadillah9838@gmail.com | Ya | Tidak | Ya | Pemecah kaca | Ya |
| hastifadillah24@sma.belajar.id | Ya | Ya | Ya | Ya | Ya |
| deahsb6@gmail.com | Ya | Ya | Ya | Pemecah kaca | Ya |
| arsiahdwicintana114@gmail.com | Ya | Ya | Ya | Ya | Ya |

| | | | | | |
|---|---|---|---|---|---|
| deasyahfirahsb@gmail.com | Ya | Ya | Ya | Ya | Ya |
| windasitorus20@gmail.com | Ya | Ya | Ya | Ya | Ya |
| rahmadanin484@gmail.com | Ya | Ya | Ya | Adanya gas untuk pemadam kebakaran | Ya |
| hafizaahasibuan03@gmail.com | Ya | Ya | Ya | Tidak | Ya |
| hnur87531@gmail.com | Ya | Tidak | Ya | Pemecah kaca | Ya |

*Table 3. Cleaning Dataset*

| No | Account name | Cleaning text | Sentiment |
|---|---|---|---|
| 0 | deasyahfira15@gmail.com | Pemecah kaca | positif |
| 1 | yunilaras2003@gmail.com | Kadang saya bisa duduk, kadang berdiri | negatif |
| 2 | diajengwulan2004@gmail.com | Pemecah kaca | positif |
| 3 | maulizakhabira@gmail.com | Pemecah kaca | positif |
| 4 | ihsantiara2127@gmail.com | Pemecah kaca | positif |

### 3.2 Pre-processing Data

Surveys are the data source used in this investigation. After cleaning, the number of data categories is labeled with sentiment (either positive or negative) based on the results of the classified test data. The accuracy level of each classification is then displayed as a percentage. Pre-processed data from 50 respondents is categorized according to their respective numbers, with 40 respondents falling into the positive category and 10 falling into the negative category. Table 3 shows the top five samples here.

Labeling is done after the dataset is loaded into Google Colab. Positive or negative labels will be applied to the analysis of the collected data. The labeling procedure is done manually. To facilitate the sentiment analysis process, data preprocessing steps are then completed. Case folding, tokenizing, normalization, filtering (Stopword Removal), and stemming are the first steps in the multi-stage data preprocessing process. The data processing step called "case folding" converts all letters to lowercase. The str.lower() function is used in the case folding stage. The tokenizing stage will remove punctuation, numbers, and spaces to split the string into tokens. This stage will also use word_tokenize() [13]. Table 4 is a sample of a tokenized review.

Next, Normalization is making everyday words or words that are still less standard into good and correct Indonesian words. Next, the standard word process is complete; the word is filtered (stopwords are removed) using Indonesian stopwords sourced from data framework filtering using the NLTK library. Finding root words and removing affixes is the last step in stemming. Tables 5 and 6 are the sample results of the stemming and filtering process.

Table 4. *Result of tokenization*

| # Tokenisasi teks<br>Tokenized = data[‘text_cleaning’]. apply(lambda x: x,split())<br><br># Menampilakn hasil tokenisasi<br>Print(tokenized.head()) | |
|---|---|
| | |
| 0 | [Pemecah, kaca] |
| 1 | [Terkadang, saya, dapat, kursi, , terkadang, be... |
| 2 | [Pemecah, kaca] |
| 3 | [Pemecah, kaca] |
| 4 | [Pemecah, kaca] |
| Name : text-cleaning, dtype : object | |

*Table 5. Result of filtering*

| | normalized_text |
|---|---|
| 0 | pemecah kaca |
| 1 | terkadang kursi terkadang berdiri halte |
| 2 | pemecah kaca |
| 3 | pemecah kaca |
| 4 | pemecah kaca |

*Table 6. Result of stemming*

| | normalized_tokenized |
|---|---|
| 0 | [pemecah, kaca] |
| 1 | [terkadang, kursi, terkadang, berdiri, halte] |
| 2 | [pemecah, kaca] |
| 3 | [pemecah, kaca] |
| 4 | [pemecah, kaca] |


*Figure 1. Wourd Cloud Negative Sentiment*


*Figure 2. Word Cloud Positive Sentiment*

### 3.3 Labeling

In the next step, we compared the amount of positive and negative review data; the sourced reviews were then combined into a new column called reviews_clean. Students gave positive responses to the use of the Trans Metro Deli Bus service. Data that is often used in reviews shows this. All positive and negative data are processed in this data visualization, which is created independently and then displayed as a word cloud, as presented in Figure 1. The most frequently seen words in the questionnaire with negative sentiment labels are shown in Figure 1. The most frequently appearing terms that resulted in negative reviews include outdated buses, crowded buses, difficulty finding a seat, inappropriate bus stops, and so on.

Meanwhile, Figure 2 visualizes the positive sentiments found in the questionnaire given. The terms often seen to generate good comments are "save money," "emergency glass breaker," and so on. After all the words are combined, the review data documents are subjected to TF-IDF weighting, which is calculated using Python words. Textual data can be converted into numeric data using TF-IDF. In addition, the TF-IDF method can improve the accuracy of the analysis. Term frequency (TF) and Inverse Document Frequency (IDF) are two concepts that are combined to determine the weight of words in the TF-IDF method [14].

### 3.4 Splitting data

The Tf-idfVectorizer(), a scikit-learn module, is used in the TF-IDF weighting calculation process. The Naïve Bayes algorithm model is then used to separate the tf-idf weighting results into training and testing data for the predictive classification process. The purpose of creating training and testing data is to separate the pre-processed data into a specific comparison. This process utilizes the sklearn library to split the training and testing data into a percentage of 75% training data and 25% testing data and data selection labels, which are independent variables. Model selection and train-test split modules simultaneously. The label column is one of the parameters used for predictive classification. The following results were obtained when user review sentiments with two positive and negative classes were classified using the Naïve Bayes algorithm method.

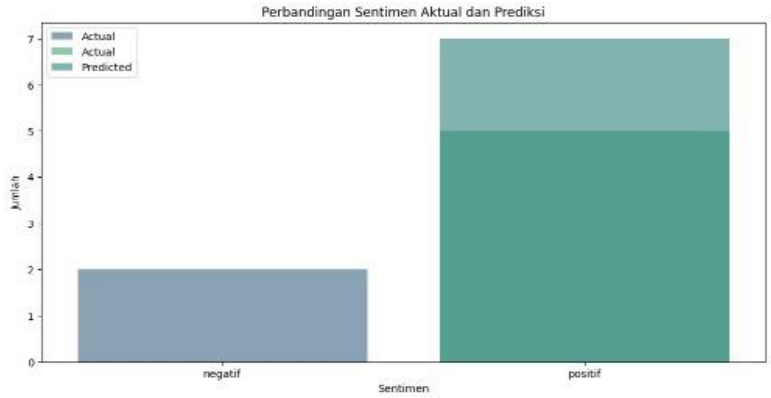


*Figure 3. Comparison of Actual and Forecasted sentiment*

*Table 7. Naïve Bayes algorithm classification results*

| Metrik | Performance | | | |
|---|---|---|---|---|
| | **precision** | **recall** | **f1-score** | **support** |
| Negative | 0.00 | 0.00 | 0.00 | 2 |
| positive | 0.71 | 1.00 | 0.83 | 5 |
| Accuracy | | | 0.71 | 7 |
| Marco avg | 0.36 | 0.50 | 0.42 | 7 |
| Weighted avg | 0.51 | 0.71 | 0.60 | 7 |

### 3.5 Implementation of Naïve Bayes Algorithm

Figure 3 shows the results of the Naïve Bayes algorithm model, which offers a relatively good accuracy of up to 71.43%. This figure shows a high level of accuracy of sentiment analysis using the Naïve Bayes classification, with accuracy results 71.43%.

## 4. CONCLUSION

The work 'hemat biaya', 'pecahan kaca', and 'nyaman' were the most frequent words in positive reviews, which are pronouns that describe the purchased product, according to the study. The word 'kedatangan bus lama' and 'tidak mendapat tempat duduk' were the most frequent negative reviews, indicating that some customers believed the product's material was flimsy. The study's Naïve Bayes results showed that people who reviewed the Trans Metro Deli Bus expressed more positive opinions, with the highest score at 71.43%.

## REFERENCES

[1] W. Putri, L. Levyda, and T. Hardiyanto, (2021), "Pengaruh Harga, Kualitas Produk, Kualitas Pelayanan, dan Promosi terhadap Keputusan Pembelian Produk," *Management & Accounting Expose*, vol. 4, no. 2, pp. 129–138.

[2] M. Khairani, F. Rakhmawati, and H. Cipta, (2022), "Penentuan Faktor-Faktor Prioritas Yang Mempengaruhi Masyarakat Dalam Penggunaan Bus Trans Metro Deli Medan Dengan Analisis Metode Analytic Hierarchy Process," *Jurnal Informatika Teknologi dan Sains*, vol. 4, no. 4, pp. 400–405, doi:10.51401/jinteks.v4i4.2137.

[3] M. I. Santoso and A. R. Dzikrillah, (2024), "Analisis Sentimen Pengguna Terhadap Kinerja Sistem Transportasi Umum Jakarta Menggunakan Algoritma Naive Bayes," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 6, pp. 3032–3043, doi:10.30865/klik.v4i6.1936.

[4] F. V. Sari and A. Wibowo, (2019), "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *Jurnal SIMETRIS*, vol. 10, no. 2, pp. 681–686.

[5] J. Ling, I. P. E. Kencana, and T. B. Oka, (2014), "Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square," *E-Jurnal Matematika*, vol. 3, no. 3, p. 92, doi:10.24843/MTK.2014.V03.I03.P070.

[6] E. P. Kondy, S. Siswanto, and N. Ilyas, (2024), "Data Balancing Approach Using Combine Sampling on Sentiment Analysis With K-Nearest Neighbor," *SISTEMASI*, vol. 13, no. 5, p. 1836, doi:10.32520/STMSI.V13I5.4013.

[7] Y. Yunefri, Y. E. Fadrial, and S. Sutejo, (2021), "Chatbot Pada Smart Cooperative Oriented Problem Menggunakan Natural Language Processing dan Naive Bayes Classifier," *INTECOMS: Journal of*

*Information Technology and Computer Science*, vol. 4, no. 2, pp. 131–140, doi:10.31539/intecoms.v4i2.2704.

[8] O. I. Gifari, M. Adha, F. Freddy, and F. F. S. Durrand, (2022), "Film Review Sentiment Analysis Using TF-IDF and Support Vector Machine," *Journal of Information Technology*, vol. 2, no. 1, pp. 36–40.

[9] E. Kurnianto and D. Febriawan, (2023), "Analisis Sentimen Perbedaan Pendapat Netizen Indonesia Terhadap Penutupan Tiktok Shop Menggunakan Algoritma Naïve Bayes," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 5, no. 2, p. 404, doi:10.30865/json.v5i2.7170.

[10] A. Aziz, (2022), "Analisis Sentimen Identifikasi Opini Terhadap Produk, Layanan dan Kebijakan Perusahaan Menggunakan Algoritma TF-IDF dan SentiStrength," *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 6, no. 1, p. 115.

[11] A. Salsabila, R. Yunita, and C. Rozikin, (2021), "Identifikasi Citra Jenis Bunga menggunakan Algoritma KNN dengan Ekstrasi Warna HSV dan Tekstur GLCM," *Technomedia Journal*, vol. 6, no. 1, pp. 124–137, doi:10.33050/tmj.v6i1.1667.

[12] N. Riska Fadhila *et al.*, (2023), "Implementasi Machine Learning Untuk Klasifikasi Narasi Informative dan Non-Informative Pada Media Sosial Twitter TMC Polda Metro Jaya Menggunakan Naïve Bayes Classifier," *Jurnal Komputer Antartika*, vol. 1, p. 2023.

[13] A. U. T. Ama, D. N. Mulya, Y. P. D. Astuti, and I. B. G. Prasadhya, (2022), "Analisis Sentimen Customer Feedback Tokopedia Menggunakan Algoritma Naïve Bayes," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, p. 50, doi:10.30865/json.v4i1.4783.

[14] N. K. Widyasanti, I. K. G. Darma Putra, and N. K. Dwi Rusjayanthi, (2018), "Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, p. 119, doi:10.24843/JIM.2018.V06.I02.P06.