# Identifying Opinions of Footwear Products in Indonesia via Sentiment Analysis

Raymond Setiawan[1*], Jimmy Tjen[2]
[1] Department of Digital Business, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
[2] Department of Informatics, Faculty of Information Technology, Universitas Widya Dharma Pontianak, Pontianak, Indonesia.
Email:[1] 22430177@widyadharma.ac.id, [2] jimmy.tjen@mathmods.eu

**Abstract.** In the era of e-commerce has significantly altered consumer interactions, this study used the Naïve Bayes Classifier to analyze sentiments from a total 1,103 data collected from Shopee. Reviews were categorized into three sentiments to provide insights for potential buyers and manufacturers. The Naïve Bayes algorithm demonstrated an overall accuracy of 87.78 precent for the dataset and 95.59 percent for self-training data. Preprocessing steps included text cleaning, normalization, and tokenization to ensure data quality. Confusion matric for three different brands revealed effective sentiment classification. Future research could extend these findings by examining additional datasets or advanced machine learning techniques to further refine sentiment classification approaches.

**Keyword:** footwear product; naïve bayes classifier; opinion mining; sentiment analysis

## Introduction

The rapid growth of e-commerce in the digital era has transformed the way customer interactions with products. This shift is particularly evident in the footwear industry, where platforms like Shopee facilitate open sharing of product reviews using comments section. By analyzing customer opinions expressed in online reviews, businesses can gain valuable insights that influence purchasing decision from the buyer's perspective. The Naïve Bayes algorithm is recognized as one of the most effective methods, and its accuracy improves when applied to larger training datasets (Chen, Hu, Hua, & Zhao, 2021).

The Naïve Bayes Classifier method is a supervised technique that can groups object by assigning class labels to records using conditional probabilities. The Bayes Theorema itself is a mathematical model based on statistics and probability. Although not something new, this algorithm remains relevant to the recent development of machine learning, especially those related to Natural Language Processing (NLP) problems. There are few research papers that be the references for this paper that utilized this methodology, see e.g (Kewsuwun & Kajornkasirat, 2022), (Yana, Santoso, & others, 2020), (Sur19), (Dey, et al., 2020), (Khomsah & others, 2020). (Kewsuwun & Kajornkasirat, 2022) This research evaluates sentiments by classifying comments into positive and negative using the Naïve Bayes Classifier, it aims to assess the sentiments and attitudes of both individuals and investors .(Yana, Santoso, & others, 2020) the dataset include around 600 records, with results analyzed through a confusion matrix to determine accuracy . (Dey, et al., 2020) The aim of this research is to compare of different machine learning techniques, specially the Support Vector Machine (SVM) and Naïve Bayes Classifier, using statistical measurement. (Khomsah & others, 2020) The objective of this study is to assess the effectiveness of Particle Swarm Optimization (PSO) in enhancing the performance of the Naïve Bayes Classifier.

The study focuses on analyzing various footwear products widely used by the public to identify the best in terms of quality, seller services, and reliability. The study aims to support potential and new customers in making informed decisions by providing objective and comprehensive evaluations of the available options in the market. The objectives of this research are twofold. First, to identify the best-quality footwear products based on critical factors such as comfort, durability, design, and prices. Second, to assess seller service reliability, which plays a crucial role in shaping the customer experience. By doing the study seeks to enhance decision-making processes for prospective buyers and provide manufacturers with feedback for continuous improvement.

The paper is organized as follows: in Section Method we describe the setup dataset used in this paper has been generated; in Section Result and Discussion we describe the dataset that has been tested using the Naïve Bayes Classifier method and illustrate the results that have been tested; in Section Conclusion we describe that what has been done in this research, show the results of the research, and the estimates of how the data can be used to other research in the feature.

## Methods

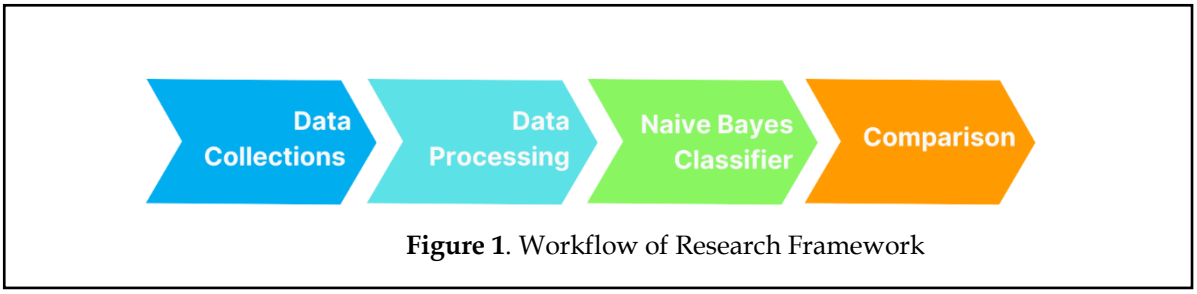


**Figure 1**. Workflow of Research Framework

*Data Collections*

The dataset were using automated web scrapping to collect all data to ensure data quality. The data were classify to 3 sentiments such as "Positive", "Negative" and "Neutral" based on reviews from comments to sentiments expressed. The total data collected is around 500 reviews data collected from several stores that sell the same footwear products and collected through Shopee. The API is utilized to gather data and conduct simulations based on reviews from the Shopee website to identify positive and negative sentiments (Sulindawaty, Laia, & Yamin, 2023).

*Data Processing*

This sections were processed using Text Analysis Toolbox to prepared the collected dataset, the processed dataset included several steps, because the data collected has not been neatly organized, so the data that has been collected is separated by comma symbols, and symbols that can't be read by the system. Therefore, is necessary to take several steps. The steps taken are as follows :

1. Text Cleaning : The non-text were removed irrelevant symbols such as emojis, symbols, and unnecessary punctuation marks to make sure the dataset consistent and changed the non-text to.
2. Normalization Data : Reviews were convert all text to lowercase to maintain uniformity dataset
3. Tokenization : The dataset were split text into individual words or array, which used to facilitates further analysis, such as ['This','products','is','amazing'].
4. Data Dictionary : The result of dataset are stored in one array, which aims to make it easier to conduct research and determine the classification of sentiments such as "Positive","Negative", and "Neutral".

*Naïve Bayes Classifier*

The algorithm that use the concept of probability for classification for sentiment analysis is known as a Naïve Bayes Classifier (Rizkya, Rianto, & Gufroni, 2023). The Naïve Bayes algorithm is a robust machine learning classification method used on probability, comparable in effectiveness to other algorithm (Syahputra, Yanris, & Irmayani, 2022).

*Comparison*

After completing the analysis of those three different footwear products, the result will be thoroughly examined and compared to identify which product demonstrates the best overall performance. This comparison will take into various factors, such as quality, comfort, price, etc, to ensure a comprehensive evaluation. The objective is to determine the product that stands out in meeting consumer needs and preferences, providing valuable insights for both potential customers and manufacturers.

## Result and Discussion

In this section, the author used a dataset of 1.103 which were used to train a whole dataset and train itself. The accuracy for whole dataset has reached 87.78 percent, and for data itself has reached 95.59 percent.

**Table 1** Accuracy Tabel

| Data Model Accuracy | |
|---|---|
| Akurasi | 87.87% |
| Akurasi2 | 95.59% |

*Confusion Matrix*

**Table 2.** Confusion Matrix Brand A

| Confusion Matrix | | Predicted | | | Total |
|---|---|---|---|---|---|
| | | Positive | Neutral | Negative | |
| Total | Positive | 315 | 0 | 1 | 316 |
| | Neutral | 11 | 30 | 0 | 41 |
| | Negative | 11 | 0 | 10 | 21 |
| Total | | 338 | 30 | 11 | 378 |

The model for Brand A indicates out of 378 data, its correctly predicted 316 positive sentiments, 41 neutral sentiments, and 21 negative sentiments. Of the 316 positive predictions, 315 were correctly identified as positive, 0 were predicted as Neutral, and 1 were predicted as Negative. Of the 41 neutral predictions, 11 were predicted as positive, 30 were correctly predicted as Neutral, and 0 were predicted as Negative. Of the 21 negative predictions, 11 were predicted as negative, 0 were predicted as Neutral, and 10 were correctly identified as Negative. The test accuracy obtained from Naïve Bayes for Brand A is 91.15 percents, and the training accuracy is 94.17 percent.

**Table 3.** Confusion Matrix Brand B

| Confusion Matrix | | Predicted | | | Total |
|---|---|---|---|---|---|
| | | Positive | Neutral | Negative | |
| Total | Positive | 315 | 0 | 1 | 316 |
| | Neutral | 11 | 30 | 0 | 41 |
| | Negative | 9 | 1 | 11 | 21 |
| Total | | 335 | 31 | 12 | 378 |

The model for Brand A indicates out of 378 data, its correctly predicted 316 positive sentiments, 41 neutral sentiments, and 21 negative sentiments. Of the 316 positive predictions, 315 were correctly identified as positive, 0 were predicted as Neutral, and 1 were predicted as Negative. Of the 41 neutral predictions, 11 were predicted as positive, 30 were correctly predicted as Neutral, and 0 were predicted as Negative. Of the 21 negative predictions, 9 were predicted as negative, 1 were predicted as Neutral, and 11 were correctly identified as Negative. The test accuracy obtained from Naïve Bayes for Brand B is 91.15 percents, and the training accuracy is 94.17 percent.

**Table 4.** Confusion Matrix Brand C

| Confusion Matrix | | Predicted | | | Total |
|---|---|---|---|---|---|
| | | Positive | Neutral | Negative | |
| Total | Positive | 315 | 0 | 1 | 316 |
| | Neutral | 11 | 30 | 0 | 41 |
| | Negative | 11 | 0 | 10 | 21 |
| Total | | 337 | 30 | 11 | 378 |

The model for Brand A indicates out of 378 data, its correctly predicted 316 positive sentiments, 41 neutral sentiments, and 21 negative sentiments. Of the 316 positive predictions, 315 were correctly identified as positive, 0 were predicted as Neutral, and 1 were predicted as Negative. Of the 41 neutral predictions, 11 were predicted as positive, 30 were correctly predicted as Neutral, and 0 were predicted as Negative. Of the 21 negative predictions, 11 were predicted as negative, 0 were predicted as Neutral, and 10 were correctly identified as Negative. The test accuracy obtained from Naïve Bayes for Brand C is 91.16 percent, and the training accuracy is 94.18 percent.

**Tabel 5.** BoW Tabel

| Word Accuracy | | Brand A | | Brand B | | Brand C | |
|---|---|---|---|---|---|---|---|
| No | Kata | Freq | Accuracy | Freq | Accuracy | Freq | Accuracy |
| 1 | Sesuai | 267 | 96.69% | 283 | 96.69% | 299 | 96.84% |
| 2 | Barang | 262 | 94.89% | 262 | 95% | 262 | 94.77% |
| 3 | Dan | 222 | 97.89% | 211 | 98.15% | 214 | 98.15% |
| 4 | Bagus | 181 | 97.09% | 184 | 97.07% | 204 | 97.36% |
| 5 | Keren | 157 | 98.83% | 167 | 98.70% | 163 | 98.89% |
| 6 | Cepat | 175 | 99.25% | 171 | 99.21% | 168 | 99.23% |
| 7 | Ori | 130 | 92.29% | 129 | 92.32% | 109 | 92.57% |
| 8 | Sepatu | 73 | 89.93% | 63 | 87.02% | 70 | 87.80% |
| 9 | Pengiriman | 127 | 95.57% | 127 | 95.14% | 125 | 95.32% |
| 10 | Nyaman | 113 | 98.37% | 107 | 97.94% | 121 | 98.5% |

The results of this study revealed significant differences in word accuracy among Brand A, Brand B, and Brand C. The Brand C gets better accuracy across most evaluated words. Brand C reached an accuracy of 99.25 percent for "cepat" and 98.84 for "sesuai", while Brand A also performed well, particularly with the same words. Brand B has slightly lower accuracy rates, particularly for word such as "sepatu". The most mentions word is "sesuai" (299 comments), and "bagus" (204 comments), highlighting the significance of there descriptors in shaping consumer perceptions of Brand C. Overall Brand C demonstrated the highest accuracy in word, and this indicate a stronger brand image or better customer experience. Then Brand A and Brand B exhibited relatively close accuracy rates, Brand A narrowly edged out in several categories, indicating that while both brands are competitive.

## Conclusion

This study a dataset collected of 1,103 reviews to evaluate the performance of sentiment analysis model using the Naïve Bayes Classifier. The model has a overall accuracy of 87.78 percent of the dataset train, and 95.59 percent of the self-training data demonstrates for the model and classifying sentiments into positive, neutral, and negative categories.

This result indicates the effectiveness of the Naïve Bayes in accurately classifying sentiment expressed. The model included confusion matrix for three different brands, which is provided data from confusion matrix

for each brand. Future research could build on these findings by exploring additional datasets or alternative machine learning techniques to further improve sentiment classification.

**References**


Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, 30.

Dey, S., Wasif, S., Tonmoy, D. S., Sultana, S., Sarkar, J., & Dey, M. (2020). A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. 217-220.

Kewsuwun, N., & Kajornkasirat, S. (2022). A sentiment analysis model of agritech startup on Facebook comments using naive Bayes classifier.

Khomsah, S., & others. (2020). Naive bayes Classifier optmization on sentiment analysis of hotel reviews. 157-168.

Ressan, M. B., & Hassan, R. F. (2022). Naive-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets. *Indonesian Journal of Electrical Engineering and Computer Science*, 375.

Rizkya, A. T., Rianto, R., & Gufroni, A. I. (2023). Implementation of the Naive Bayes Classifier for Sentiment Analysis of the Shopee E-Commerce Application Review Data on the Google Play Store. *Internastional Journal of Applied Information Systems and Informatics (JAISI)*, 31-37.

Sulindawaty, S., Laia, E., & Yamin, M. (2023). Penerapan Algoritma Naive Bayes dalam Menganalisis Setnimen pada Review Pengguna E-Commerce. *KLIK: Kajian Ilmiah Informatika dan Komputer*, 305-316.

Syahputra, R., Yanris, G. J., & Irmayani, D. (2022). Svm and naive bayes algorithm comparison for user sentiment analysis on twitter. *Sinkron: Jurnal dan penelitian teknik informatika*, 674.

Yana, A., Santoso, T., & others. (2020). Sentiment analysis of facebook comments on indonesian presidential candidates using the naive bayes method. 12-24.