Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Improving Sentiment Analysis Performance of Tokopedia Reviews Using Principal Component Analysis and Naïve Bayes Algorithm

Anjar Ayuning Lestari[1]*, Ahmad Faqih[2], Gifthera Dwilestari[3]*

[1,2]*Informatics Engineering, STMIK IKMI Cirebon*
[3]*Information System, STMIK IKMI Cirebon*
anjarayuningl@gmail.com[1*]

Abstract

Tokopedia one of Indonesia's largest e-commerce platforms, offers a wide range of products with diverse customer reviews. These reviews reflect consumer opinions and provide valuable insights for service improvement and marketing strategies. Sentiment analysis is crucial for understanding customer perceptions, but processing large-scale, high-dimensional text data remains a challenge, impacting model efficiency and accuracy. This research uses Principal Component Analysis (PCA) to reduce data dimensionality without losing important information for sentiment classification. The study begins by collecting Tokopedia product reviews and preprocessing the text, including data cleaning, tokenization, stopword removal, and stemming. The reviews are then converted into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) method. A Gaussian Naïve Bayes model is employed to classify sentiment into three categories: positive, neutral, and negative. The results demonstrate that PCA significantly improves model accuracy from 63.13% to 70.47%, with gains in precision (71.85%), recall (70.47%), and F1-score (71.06%). This research contributes to enhancing sentiment analysis techniques using PCA for Tokopedia reviews and offers a valuable approach that can be applied to other e-commerce platforms.

Keywords: *Principal Component Analysis(PCA),Naïve Bayes,sentiment analysis,Tokopedia,TF-IDF,accuracy,e-commerce*

1. Introduction

The rapid growth of e-commerce users in Indonesia, especially on the Tokopedia platform, has resulted in a large transaction volume, which requires efficient data analysis, particularly related to user reviews. Sentiment analysis of these reviews can provide valuable insights for companies to enhance user experience and services. Although the Naïve Bayes algorithm has been proven effective in text classification, its performance can be improved by integrating Principal Component Analysis (PCA) to reduce the data dimensionality without sacrificing accuracy.

Previous studies, such as those by [1] and [2], identified challenges in applying Naïve Bayes to large and complex datasets. They emphasized the importance of feature optimization techniques, such as TF-IDF, N-gram, and PCA, to improve classification performance. This study aims to combine PCA and Naïve Bayes to optimize sentiment analysis on Tokopedia reviews, addressing the challenges posed by the large data volume. PCA will reduce dimensionality by eliminating irrelevant or redundant features, allowing Naïve Bayes to work more efficiently.

The method used in this study involves applying PCA to reduce data dimensionality, followed by sentiment classification using Naïve Bayes. This approach will be evaluated through experiments with Tokopedia review datasets, comparing the performance of traditional methods with the proposed one. Evaluation will be carried out by measuring metrics such as accuracy, precision, recall, F1-score, and computation time.

This research can contribute to the development of more efficient sentiment analysis techniques for large-scale data. The integration of PCA with Naïve Bayes is expected to improve accuracy and efficiency in sentiment classification, and can be applied to various e-commerce platforms as well as other industries managing large data, to support business decisions and enhance user services.

2. Research Methods

This research uses the KDD (Knowledge Discovery in Database) technique due to its advantages in identifying organized patterns from complex datasets and making the data easier to understand. The KDD (Knowledge Discovery in Databases) method is a process of knowledge discovery in databases that applies scientific methods to data mining. Knowledge Discovery in Databases (KDD) is a process that involves studying data, developing models, and deriving algorithms[3]. The goal of KDD (Knowledge Discovery in Database) and data mining is to extract hidden information from large datasets. The stages are shown in the figure below.

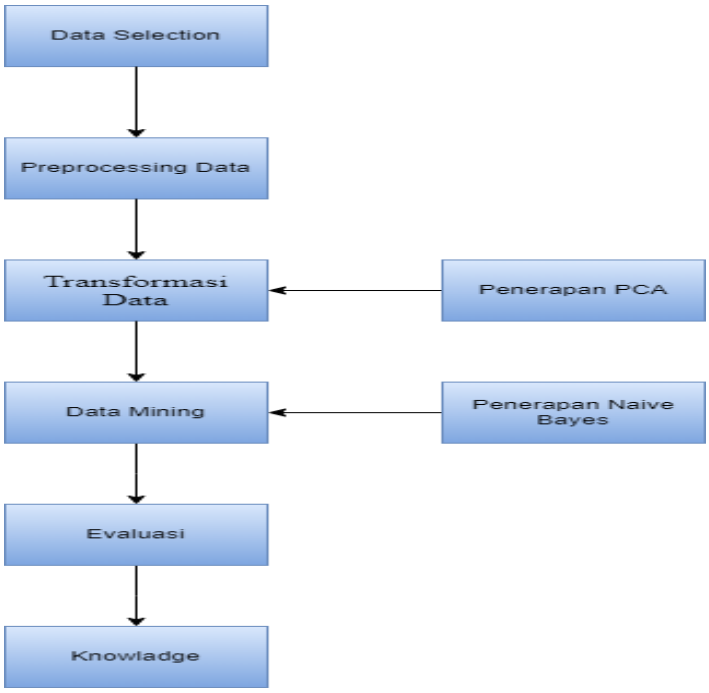

Fig. 1: Research Method Stages

The following is an explanation of the stages in Figure 1:

2.1. Data Selection

The dataset used in this research was obtained from user reviews of the Tokopedia app on the Google Play Store[4]. The collected data includes reviewId, username, content, score, and other attributes. This data is used for sentiment classification (Negative, Positive, Neutral) using the Naïve Bayes algorithm and PCA. The results are shown in Table.1:

Table 1: Result Data Selection

…..	userName	content	score	….
…..	Pengguna Google	Untuk pengembang, tolong yang sistem "BEBAS ONGKIR / Rekomendasi" bisa memilih kurirnya, jangan otomatis, tiap daerah ga semua ada kurirnya (Ada tapi beda kota), alhasil paket stuck / mau ngambil ke tempat juga jauh. Yang lain udah OK, Toko Online Favorit saya harus lebih baik lagi. Terima kasih 🙏	5	….

2.2. Data Preprocessing

Preprocessing is crucial for producing high-quality data and improving the accuracy of the Naïve Bayes algorithm's performance[5]. During the preprocessing stage, data is collected and prepared for further analysis. This step is performed to ensure that the data can be effectively used in the machine learning model by simplifying and organizing it. Below is an explanation of several preprocessing steps.

2.2.1. Cleaning

In this stage, irrelevant elements such as HTML tags, emojis, numbers, and punctuation are removed from the review text[6]. The goal of this process is to produce cleaner and more consistent text, making it easier for subsequent analysis. The cleaning process is carried out using specialized functions that automatically remove these elements, with the expectation of improving the accuracy and clarity of sentiment analysis results. The results are shown in Table.2:

Table 2: Result Cleaning

cleaning
Untuk pengembang tolong yang sistem BEBAS ONGKIR Rekomendasi bisa memilih kurirnya jangan otomatis tiap daerah ga semua ada kurirnya Ada tapi beda kota alhasil paket stuck mau ngambil ke tempat juga jauh Yang lain udah OK Toko Online Favorit saya harus lebih baik lagi Terima kasih

2.2.2. Case Folding

After the cleaning process, the next step is case folding, which involves converting all text to lowercase. The purpose of this step is to unify word variations caused by capitalization differences, thereby improving data consistency[7]. With case folding, words like 'Data', 'DATA', and 'data' will be treated as the same form, which simplifies the analysis process and enhances the accuracy of the results. The results are shown in Table.3:

Table 3: Result Case Folding

case_folding
untuk pengembang tolong yang sistem bebas ongkir rekomendasi bisa memilih kurirnya jangan otomatis tiap daerah ga semua ada kurirnya ada tapi beda kota alhasil paket stuck mau ngambil ke tempat juga jauh yang lain udah ok toko online favorit saya harus lebih baik lagi terima kasih

2.2.3. Tokenization

The next step is tokenization, which is the process of breaking down text into smaller units, known as tokens, which are typically individual words. In this research, tokenization is applied to the text that has gone through the case folding stage to separate it into individual words. This process enables deeper analysis by identifying each word as a standalone unit.The results are shown in Table.4:

Table 4: Result Tokenization

tokenization
['untuk', 'pengembang', 'tolong', 'yang', 'sistem', 'bebas', 'ongkir', 'rekomendasi', 'bisa', 'memilih', 'kurirnya', 'jangan', 'otomatis', 'tiap', 'daerah', 'ga', 'semua', 'ada', 'kurirnya', 'ada', 'tapi', 'beda', 'kota', 'alhasil', 'paket', 'stuck', 'mau', 'ngambil', 'ke', 'tempat', 'juga', 'jauh', 'yang', 'lain', 'udah', 'ok', 'toko', 'online', 'favorit', 'saya', 'harus', 'lebih', 'baik', 'lagi', 'terima', 'kasih']

2.2.4. Stopword Removal

Stopword removal is the step of removing words that frequently appear but carry little meaning in text analysis, such as 'yang', 'di', and 'dari'. This process is performed to ensure that the analysis focuses on more significant words. In this research, the Indonesian stopwords list from NLTK is used. The results are shown in Table.5:

Table 5: Result Stopword Removal

stopword_removal
['pengembang', 'tolong', 'sistem', 'bebas', 'ongkir', 'rekomendasi', 'memilih', 'kurirnya', 'otomatis', 'daerah', 'ga', 'kurirnya', 'beda', 'kota', 'alhasil', 'paket', 'stuck', 'ngambil', 'udah', 'ok', 'toko', 'online', 'favorit', 'terima', 'kasih']

2.2.5. Normalization

The goal is to replace non-standard words or slang with their formal language equivalents, making the text ready for analysis. In this research, normalization is performed using a slang dictionary that contains pairs of slang words and their standard language equivalents, stored in JSON format, to replace slang words with their formal counterparts. The results are shown in Table.6:

Table 6: Result Normalization

normalization
pengembang tolong sistem bebas ongkos kirim rekomendasi memilih kurirnya otomatis daerah ga kurirnya beda kota alhasil paket stuck mengambil sudah ok toko online favorit terima kasih

2.2.6. Steamming

Stemming is the process of converting words to their base or root form by removing affixes. This process simplifies word variations that have similar meanings, allowing the analysis to focus on the core meaning. In this research, stemming is performed using the Sastrawi library for the Indonesian language. The results are shown in Table.7:

Table 7: Result Stemming

stemming
kembang tolong sistem bebas ongkos kirim rekomendasi pilih kurir otomatis daerah ga kurir beda kota alhasil paket stuck ambil sudah ok toko online favorit terima kasih

2.2.7. Labelling

Labeling is the process of identifying and categorizing emotions or sentiments in text into classes such as positive, neutral, or negative. In this research, sentiment analysis is performed using the nlptown/bert-base-multilingual-uncased-sentiment model from Hugging Face Transformers, which is capable of analyzing text in multiple languages, including Indonesian. The results are shown in Table.8:

Table 8: Result Labelling

stemming	label
kembang tolong sistem bebas ongkos kirim rekomendasi pilih kurir otomatis daerah ga kurir beda kota alhasil paket stuck ambil sudah ok toko online favorit terima kasih	Netral
…..	…..
aplikasi sial x beli pakai voucher dibatalin sistem saja nyoba yang tiga mas dibatalin sistem emg anjg iya tokped bagus msi bagus shopee	Negatif

2.3. Data Transformation

At this stage, the previously processed text data is then transformed using two main methods: TF-IDF (Term Frequency-Inverse Document Frequency) and PCA (Principal Component Analysis). TF-IDF is used to extract text features based on the relative frequency of words in documents, generating a numerical representation ready for analysis. After that, PCA is applied to reduce the dimensionality of the TF-IDF data, decreasing the number of features without losing important information.

2.3.1. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF It is applied to extract features from the text that has undergone the stemming process. This method calculates the frequency of word occurrences in documents while considering how often the word appears across the entire collection of documents[8]. The process begins by ensuring that the stemming and label columns are present, and filling in any missing values in the stemming column with empty strings to avoid errors during the process. Below are the results of the TF-IDF:

Table 9: Result TF-IDF

Document ID	Term ID	*TF-IDF* Weight
0	7921	0.1295
0	5530	0.1868
0	2375	0.3366
0	7751	0.1697
0	3537	0.1460
...	...	...
4998	6746	0.4783
4999	7247	0.2189
4999	8594	0.1023
4999	4900	0.3380

2.3.2. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is applied to reduce the dimensionality of the data, with the aim of preserving the main information and simplifying complexity[9]. In this research, PCA is used to reduce the number of features in the TF-IDF data while retaining 50% of the variance from the original data. PCA helps reduce redundancy and improve efficiency in the analysis.

2.4. Data Mining

At this stage, the Naive Bayes model is applied to classify the processed text data in two conditions. First, the model is used on the TF-IDF data without dimensionality reduction, and second, on the data that has been dimensionally reduced using PCA. This approach aims to evaluate the impact of dimensionality reduction on the model's performance and test how much PCA can simplify the data without compromising classification accuracy.

2.4.1. Naïve Bayes Before PCA Application

At this stage, the Gaussian Naive Bayes model is initialized without applying PCA. The training and test data are converted to a dense format using the toarray() method so they can be processed by the model. The model is then trained using the training data to identify patterns, and after the training process is completed, it is used to predict the test data. The prediction results are evaluated to measure the model's performance without dimensionality reduction.

2.4.2. Naïve Bayes After PCA Application

At this stage, the Gaussian Naive Bayes model is initialized using data that has undergone PCA transformation. The dimensionality-reduced training data is used to train the model, which is then used to make predictions on the test data, also transformed using PCA. After the prediction process, the model's performance is evaluated by measuring Accuracy, Precision, Recall, and F1 Score.

2.5. Evaluation

At this stage, the trained model is evaluated to measure its performance in classifying the data. The evaluation is carried out by calculating accuracy and analyzing other metrics, such as Precision, Recall, and F1-score. These metrics are used to assess the model's ability to predict the test data, and the evaluation results provide an overview of how well the model achieves the classification objectives.

3. Result and Discussion

Results In this study, converts review text into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency) and demonstrates that the combination of TF-IDF with Naïve Bayes is effective for sentiment classification (negative, neutral, and positive) with satisfactory accuracy. TF-IDF plays an important role in identifying meaningful words in specific contexts, improving prediction accuracy by giving more weight to words that are common in a particular sentiment category but rare in others. These results align with the study by [10] which also applied the combination of TF-IDF and Naïve Bayes for sentiment analysis on beauty products, showing similar findings regarding the enhancement of model performance.

The research by [11] using SVM and TF-IDF shows that the TF-IDF weighting also improves sentiment classification accuracy in margin-based models. In addition, this study also applies PCA (Principal Component Analysis) to Naïve Bayes to reduce the dimensionality of the data, retaining 50% of the variance from the original data. The results are consistent with previous studies, such as [12], which showed that

dimensionality reduction using PCA can improve classification model accuracy, as well as the study by [13], which emphasized that PCA is effective in simplifying data while preserving most of the important information.

In addition, the results of this study Principal Component Analysis (PCA) was applied to the Gaussian Naive Bayes model to reduce the dimensionality of the data with the aim of improving computational efficiency and classification accuracy. The results show that the use of PCA increased the model's accuracy from 63.13% to 70.47%, representing an improvement of 7.33%.

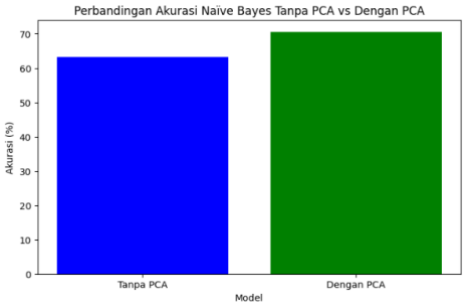

Fig. 2: Comparison Diagram of Model Accuracy

Table 10: Evaluation Matrix of Naïve Bayes Accuracy Without PCA & With PCA

Evaluation Metrics	Value Without PCA (%)	Value With PCA (%)
Accuracy	63.13	70.47
Precision	68.12	71.85
Recall	63.13	70.47
F1 Score	65.23	71.06

From Table 10, it can be seen that the application of PCA to the Naive Bayes model successfully improved all evaluation metrics. Accuracy increased from 63.13% to 70.47%, Precision from 68.12% to 71.85%, Recall from 63.13% to 70.47%, and F1 Score from 65.23% to 71.06%. This improvement indicates that PCA is effective in enhancing the performance of the Naive Bayes model. This is consistent with the research by [14], which showed that the application of Weighted Principal Component Analysis (W-PCA) successfully improved classification accuracy on high-dimensional datasets, such as Pap Smear images, from 67.45% to 87.24%. By using W-PCA, computation time was also reduced without compromising model precision. Another study by [15] also supports these findings, where the use of PCA in classifying KIP scholarship recipients' data increased Naïve Bayes accuracy to 85.19%, higher than the 83.33% obtained without PCA.

4. Conclusion

The results of this study show that the model's accuracy increased from 63.13% to 70.47%, representing an improvement of 7.33%. Additionally, other evaluation metrics also showed improvement, with Precision rising from 68.12% to 71.85%, Recall increasing from 63.13% to 70.47%, and the F1-score increasing from 65.23% to 71.06%. This improvement demonstrates that PCA is effective in reducing the dimensionality of the data without losing important information.

References

[1] M. Ala'raj, M. Majdalawieh, and M. F. Abbod, "Improving binary classification using filtering based on k-NN proximity graphs," *J. Big Data*, vol. 7, no. 1, p. 15, Dec. 2020, doi: 10.1186/s40537-020-00297-7.

[2] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 1, p. 55, 2021, doi: 10.22146/ijccs.60718.

[3] A. Y. Simanjuntak, I. S. S. Simatupang, and Anita, "Implementasi Data Mining Menggunakan Metode Naïve Bayes Classifier Untuk Data Kenaikan Pangkat Dinas," *J. Sci. Soc. Res.*, vol. 4307, no. 1, pp. 85–91, 2022.

[4] H. P. Doloksaribu and Y. T. Samuel, "Komparasi Algoritma Data Mining Untuk Analisis Sentimen Aplikasi Pedulilindungi," *J. Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform.*, vol. 16, no. 1, pp. 1–11, 2022, doi: 10.47111/jti.v16i1.3747.

[5] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Comparison of Naïve Bayes and Support Vector Machine Methods in Twitter Sentiment Analysis," *Smatika J.*, vol. 10, no. 02, pp. 71–76, 2020.

[6] M. Y. Putra and D. I. Putri, "Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI," *J. Tekno Kompak*, vol. 16, no. 2, p. 176, 2022, doi: 10.33365/jtk.v16i2.2002.

[7] J. Supriyanto, D. Alita, and A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, pp. 74–80, 2023, doi: 10.33365/jatika.v4i1.2468.

[8] A. Averina, H. Hadi, and J. Siswantoro, "Analisis Sentimen Multi-Kelas Untuk Film Berbasis Teks Ulasan Menggunakan Model Regresi Logistik," *Teknika*, vol. 11, no. 2, pp. 123–128, 2022, doi: 10.34148/teknika.v11i2.461.

[9] K. Astoni and M. Haris, "Analisis Penerapan Principal Component Analysis (Pca) Pada Deteksi Kecurangan Kartu Kredit Menggunakan Random Forest an Analysis of Principal Component Analysis Implementation on Credit Card Fraud Detection Using Random Forest," *J. Elektro Telekomun. Terap.*, vol. 9, no. 1, pp. 1152–1161, 2022, [Online]. Available: https://doi.org/10.25124/jett.v9i1.5019

[10] M. H. Wicaksono, M. D. Purbolaksono, and S. Al Faraby, "Perbandingan Algoritma Machine Learning untuk Analisis Sentimen Berbasis Aspek pada Review Female Daily," *eProceedings Eng.*, vol. 10, no. 3, pp. 3591–3600, 2023.

[11] E. Hokijuliandy, H. Napitupulu, and F. Firdaniza, "Analisis Sentimen Menggunakan Metode Klasifikasi Support Vector Machine (SVM) dan Seleksi Fitur Chi-Square," *SisInfo J. Sist. Inf. dan Inform.*, vol. 5, no. 2, pp. 40–49, 2023, doi: 10.37278/sisinfo.v5i2.670.

[12] A. Dinanti and J. Purwadi, "Analisis Performa Algoritma K-Nearest Neighbor dan Reduksi Dimensi Menggunakan Principal Component Analysis," *Jambura J. Math.*, vol. 5, no. 1, pp. 155–165, Feb. 2023, doi: 10.34312/jjom.v5i1.17098.

[13] F. Badri and S. U. R. Sari, "Penerapan Metode Principal Component Analysis (PCA) Untuk Identifikasi Faktor-Faktor yang Mempengaruhi Sikap Mahasiswa Memilih Melanjutkan Studi ke Kota Malang," *Build. Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 426–431, Dec. 2021, doi: 10.47065/bits.v3i3.1139.

[14] Y. N. Dewi, H. Rianto, D. Riana, and J. Siregar, "Integrasi Metode Sample Bootstrapping Dan Weighted Principal Component Analisys (PCA) Untuk Meningkatkan Performa Naïve Bayes Pada Citra Tunggal Papsmear," *Inti Nusa Mandiri*, vol. 14, no. 2, pp. 133–138, 2020, doi:

https://doi.org/10.33480/inti.v14i2.1103 VOL.

[15] B. Hirwono and Suhirman, "Classification of Indonesian Smart Card Scholarship Recipients with Principal Component Analysis using the Naive Bayes and Decision Tree Methods Case Study: Stie Pariwisata API Yogyakarta," *Int. J. Comput. Appl.*, vol. 186, no. 5, pp. 13–21, 2024, doi: 10.5120/ijca2024923375.