Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 2, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i2.10074

Sentiment Analysis of Shopee Application Reviews Using Multinomial Naïve Bayes and Bigram Feature Extraction

Viktor Wahyu Nugroho[1],L.Budi Handoko[2]
[1,2]*Informatics Engineering Study Program, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia*
[1]111202214176@mhs.dinus.ac.id, [2]handoko@dsn.dinus.ac.id

Info Artikel

Riwayat Artikel:

Received 2026-01-02
Revised 2026-04-10
Accepted 2026-05-14

Abstract *– The rapid growth of e-commerce in Indonesia has made user reviews a critical source of feedback, yet discrepancies between star ratings and actual sentiment often mislead businesses. This study employs the Multinomial Naïve Bayes algorithm to analyze sentiment in 25,000 Shopee application reviews collected via web scraping. The research utilizes TF-IDF for feature extraction and Bigram analysis to capture contextual meaning, addressing the challenge of imbalanced data (82% positive, 18% negative). The objective is to accurately classify user sentiment into positive and negative categories to provide actionable insights beyond numerical ratings. The model achieved a classification accuracy of 91.96%, with a high Recall of 77% for the minority negative class, ensuring effective identification of user complaints. Bigram analysis revealed that "delivery speed" is the primary driver for both satisfaction and dissatisfaction. The study confirms that Naïve Bayes is a robust and scalable solution for large-scale sentiment analysis in the Indonesian e-commerce context, offering a reliable tool for business intelligence. These findings emphasize the importance of text-based analysis over traditional rating systems, helping developers prioritize service enhancements.*

Keywords: *Bigram Analysis; E-commerce; Naïve Bayes; Sentiment Analysis; Shopee*

Corresponding Author:
Viktor Wahyu Nugroho
Email:
111202214176@mhs.dinus.ac.id

Abstrak *– Pesatnya pertumbuhan e-commerce di Indonesia menjadikan ulasan pengguna sebagai sumber umpan balik yang kritis, namun ketidaksesuaian antara peringkat bintang dan sentimen aktual sering kali menyesatkan pelaku bisnis. Penelitian ini menerapkan algoritma Multinomial Naïve Bayes untuk menganalisis sentimen pada 25.000 ulasan aplikasi Shopee yang dikumpulkan melalui web scraping. Penelitian ini memanfaatkan TF-IDF untuk ekstraksi fitur dan analisis Bigram untuk menangkap makna kontekstual, serta mengatasi tantangan data yang tidak seimbang (82% positif, 18% negatif). Tujuannya adalah untuk mengklasifikasikan sentimen pengguna secara akurat ke dalam kategori positif dan negatif guna memberikan wawasan yang dapat ditindaklanjuti di luar sekadar peringkat numerik. Model ini mencapai akurasi klasifikasi sebesar 91,96%, dengan Recall yang tinggi sebesar 77% untuk kelas negatif minoritas, memastikan identifikasi keluhan pengguna yang efektif. Analisis Bigram mengungkapkan bahwa "kecepatan pengiriman" adalah faktor utama pendorong kepuasan maupun ketidakpuasan. Studi ini mengonfirmasi bahwa Naïve Bayes adalah solusi yang tangguh dan dapat diskalakan untuk analisis sentimen skala besar dalam konteks e-commerce Indonesia, menawarkan alat yang andal untuk intelijen bisnis. Temuan ini menekankan pentingnya analisis berbasis teks dibandingkan sistem peringkat tradisional, yang bisa membantu pengembang memprioritaskan peningkatan layanan.*

Kata Kunci: *Analisis Bigram; Analisis Sentimen; E-commerce; Naïve Bayes; Shopee*

I. INTRODUCTION

In today's digital era, the development of internet infrastructure across Indonesia has fundamentally changed how consumers interact with businesses and their purchasing behavior. E-commerce has now become a rising platform for retailers, facilitating transactions between sellers and buyers and capitalizing on Indonesia's rapid e-commerce growth.[1], [2], [3] And the rapidly expanding digital economy. Applications such as Shopee, Tokopedia, and Lazada are emerging as dominant forces in the Indonesian market. These platforms have provided millions of users with convenient, accessible shopping experiences, with many Indonesians choosing Shopee as their preferred e-commerce platform.[1]. With the rapid growth in e-commerce adoption, developers face a critical challenge: understanding user experience, satisfaction, and areas for improvement. User reviews are a key metric for understanding consumer feedback, providing deeper insights into consumers' experiences. [4], [5]. However, highly rated apps with both positive and negative reviews do not necessarily indicate that a good user experience or satisfaction has been achieved. Discrepancies between numerical ratings and actual written reviews, where users give a high star rating but express frustration in their comments, can significantly lead to inconsistencies and a lack of accuracy for assessing application performance and user satisfaction. [4] And could lead to confusion for users seeking feedback.

Sentiment analysis, also known as opinion mining, is a computational technique for detecting and categorizing opinions, feelings, and emotions in text data. It also cleans data by removing irrelevant words and symbols and transforms it from qualitative to quantitative form data. [6], [7], [8] This approach enables businesses to gain a

comprehensive understanding of user feedback without relying solely on numerical ratings, which often fail to reflect the overall user experience accurately. Sentiment analysis can also classify the polarity of text to determine whether a review or written opinion is positive or negative [6]. The information obtained from this approach and analyzed, such as the proportion of the Positive and negative reviews, as well as the most frequent topics and opinions mentioned, can provide valuable insights for e-commerce companies. This information can help them improve services and product quality, refine marketing strategies, and gain a deeper understanding of customers' needs. [9]

The Naïve Bayes Algorithm is a classification method that estimates probabilities by summing values from a dataset. [6], [10], [11]. The Naïve Bayes algorithm, alongside its comparative models, has recently proven highly effective in analyzing public sentiment towards e-commerce usage [12], [13], as well as categorizing opinions across various Indonesian social media platforms and Q&A forums [14], [15], [16]. study uses Naïve Bayes due to its simplicity and speed [6]. Although simple, this algorithm often achieves performance comparable to other classification methods. In this study, the Naïve Bayes algorithm is used to process and analyze large-scale Shopee app review data from the Google Play Store [17]. The reviews undergo preprocessing, including cleaning and normalization, after which the Naïve Bayes classifier categorizes each review into sentiment classes such as Positive and Negative. Its inherent simplicity, combined with computational efficiency and demonstrated accuracy in sentiment classification, makes it well-suited for analyzing large-scale review datasets. [2], [18], [19]

Similar research has been conducted on various e-commerce applications using the Naïve Bayes algorithm with different data-splitting conditions. Three train-test configurations were evaluated across the platforms Shopee, Tokopedia, and Lazada, using Shopee reviews: 80% of the data for training and 20% for testing. This test yielded an accuracy of 92%, a precision of 92.13%, a recall of 98.8%, and an F1-score of 95.35%.[6] For Tokopedia, the best performance was achieved with a 60:40 train-test split, resulting in 83.5% accuracy, 82.58% precision, 91.6% recall, and an F1-score of 91.6%. Lazada also performed best under the same 60:40 split, with 79.5% accuracy, 79.4% precision, 100% recall, and an F1-score of 88.52%.[6] These findings indicate that Naïve Bayes exhibits strong, comparable performance across various marketplaces. However, optimizing the model's train–test setup is essential, as performance metrics can be sensitive to the specific platform and its inherent data distribution. A more detailed analysis of Naïve Bayes on Shopee reviews further demonstrates how class-level performance can differ across sentiment categories. In this study, using a two-class system, Positive and Negative Class, 18 reviews were predicted as negative. Of these, only 12 of 18 were negative, while 6 were misclassified, resulting in false positives. This resulted in relatively low precision (92.88%), despite high recall (97.44%).[9] For the positive class, 76 of 78 positive reviews were correctly identified, with only 2 misclassified instances (false negatives), yielding recall and precision of 97.44%. Overall, the Naïve Bayes classifier achieved a test accuracy of 88% on Shopee reviews, lower than its 97% training accuracy, suggesting some overfitting but still demonstrating strong generalization in practice. From a sentiment distribution perspective, Shopee reviews in this experiment were predominantly positive at 74.6% (746 out of 1000 reviews), followed by 21.2% negative and only 4.2% neutral, indicating most users expressed satisfaction [9].

To address the limitations observed in prior literature, the novelty of this research lies in the combination of Bigram feature extraction to capture contextual meaning, the utilization of a massive dataset comprising 25,000 reviews, and a specific focus on handling imbalanced data robustly. Unlike the study in [6] which relied on a limited dataset of 500 reviews, and [9] which faced precision challenges on minority classes, our approach ensures that critical negative feedback is accurately identified despite the vast dominance of positive ratings. Therefore, this research aims to conduct a systematic sentiment analysis of customer reviews of Shopee on the Google Play Store using the Naïve Bayes algorithm. By analyzing review data, the study seeks to classify user sentiment into distinct categories, providing comprehensive insights into the application's user experience. Additionally, this study contributes to a broader understanding of Naïve Bayes in analyzing unstructured text within the Indonesian language context.

II. METHOD

This research employs a systematic approach to sentiment analysis using the Knowledge Discovery in Databases (KDD) framework, which consists of sequential stages designed to transform raw review data into actionable insights. The methodology encompasses data collection, labeling, preprocessing, feature extraction, model training, and evaluation. Each stage is essential to ensure data quality and model reliability in classifying user sentiments toward the Shopee application.[6]

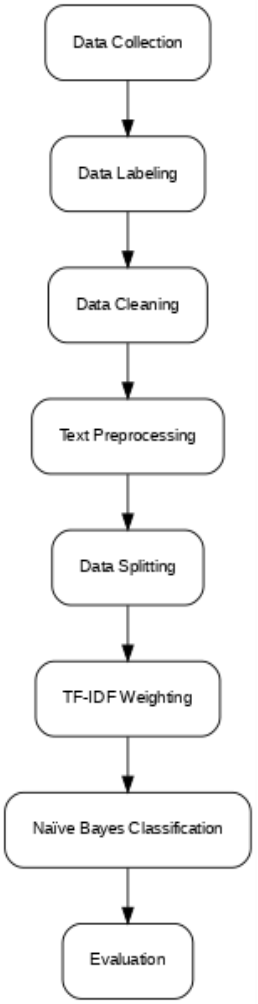

Figure 1. Research Methodology Pipeline for Sentiment Classification

The data used in this study consists of user reviews of the Shopee application obtained from the Google Play Store. The dataset comprises textual review content accompanied by metadata, including review identifiers, usernames, star ratings, and timestamps. Only the review text and its associated sentiment label are utilised as primary variables in the sentiment analysis task, whereas other attributes serve as supporting information for filtering and validation. The reviews are written in Indonesian and reflect real user experiences, including both functional and service-related aspects of the application. This dataset is highly relevant to the research objective because it directly represents users' subjective perceptions, enabling a detailed analysis of positive, negative, and neutral sentiments toward Shopee. By focusing on naturally occurring user-generated content, the study captures authentic feedback that supports the problem statement regarding the gap between numeric ratings and textual opinions.

The initial phase involves collecting user review data from the Google Play Store via web scraping. The scraping process uses the google-play-scraper library in Python to automatically extract review text, ratings, timestamps, and user information from the Shopee application page. This method is preferred over manual collection because it efficiently handles large volumes of data and captures real-time user feedback. The scraping process targets reviews posted between September 24 and October 14, 2025, to ensure data relevance and recency. The extracted data includes attributes such as reviewId, userName, content (review text), score (star rating), and at (review date). Among these attributes, the content field containing the actual review text serves as the primary data source for sentiment analysis.

reviewId	userName	content	score	at
R-1001	Aulia P.	Barang sesuai deskripsi, pengiriman cepat.	5	2024-07-12 10:15
R-1002	Dimas W.	Kualitas oke tapi packing kurang rapi.	4	2024-07-13 08:42
R-1003	Sari L.	Ukuran tidak sesuai, warna agak beda dari foto.	3	2024-07-14 17:05
R-1004	Budi K.	Pengiriman lambat, tapi penjual responsif saat ditanya.	2	2024-07-15 12:33
R-1005	Clara M.	Barang cacat, sudah ajukan retur.	1	2024-07-16 09:20

Figure 2. Sample of Raw Scraped Reviews (reviewId, userName, content, score, at)

After data collection, the reviews undergo automated sentiment labelling based on the user's star rating (score). This approach is chosen to handle the large volume of data efficiently and to eliminate the subjectivity associated with manual annotation.Labelling criteria:

- **Positive:** Reviews with a star rating of 4 or 5. These ratings typically indicate user satisfaction and favourable experiences**.**
- **Negative:** Reviews with a star rating of 1, 2, or 3. These ratings generally reflect dissatisfaction, complaints, or critical feedback.

This rating-based labelling method aligns with standard practices in sentiment analysis for e-commerce. While rating-based labeling has the potential for bias in edge cases (e.g., sarcastic reviews with high ratings), it provides a highly objective and scalable ground truth for large datasets, eliminating the subjectivity and inconsistency often associated with manual human annotation. The resulting dataset contains two distinct classes: Positive and Negative, which are then used for supervised learning.

To ensure consistency and reliability, a subset of reviews is cross-validated by multiple annotators, and disagreements are resolved through consensus. Each annotator applies the positive/negative criteria consistently across all assigned reviews, recording the label in a structured format (e.g., reviewId, text, label, annotator) that accompanies the review text.

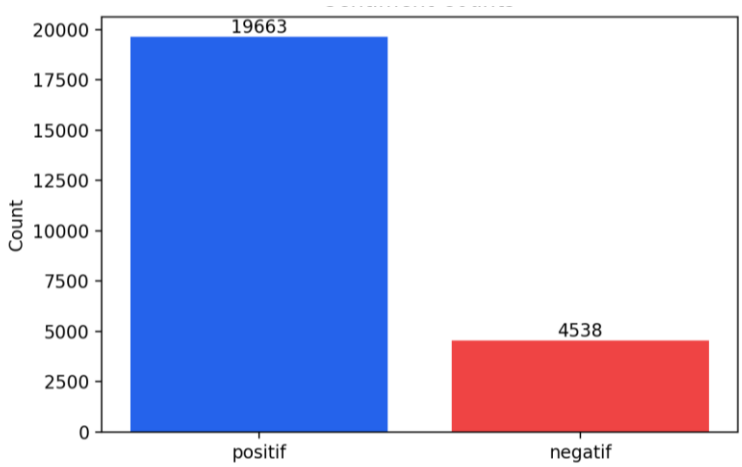

Figure 3. Distribution of Sentiment Labels (positif vs negatif)

Data cleaning addresses data quality and completeness issues in user-generated content by identifying and handling missing values, duplicate entries, and irrelevant records that could harm model performance. Missing values-especially in the review text (text_stemindo) field-are managed with an ignore-tuple (listwise deletion) strategy that removes any record with critical missing data. This is appropriate when missingness is small, and the remaining data suffice for training. Incomplete reviews that lack essential textual information are excluded to prevent noise and maintain dataset consistency for subsequent processing. In this pipeline, rows missing text_stemindo or label are dropped (listwise deletion), and the labels are restricted to the two classes (Positive, Negative). Unlike methods that down-sample the majority class, this study preserves the original imbalance (approximately 82% positive and 18% negative) to ensure that the model learns to operate in a realistic environment in which positive feedback typically outweighs negative complaints.

Text preprocessing transforms raw, unstructured reviews and standardises review text before feature extraction and classification. The dataset already contains a preprocessed text field (text_stemindo) that is used for all

embeddings and classifiers. This stage is crucial for improving model accuracy by reducing noise and normalising linguistic variations commonly found in user-generated content. The preprocessing pipeline consists of three sequential operations: case folding, stopword removal, and stemming.

Case Folding converts all characters to lowercase, so variants such as "Bagus" and "BAGUS" map to the same token, thereby reducing the vocabulary and aiding pattern recognition. Stopword removal filters high-frequency, low-information function words (e.g., “yang”, “dan”, “di”, “dengan”, “untuk”, “adalah”) to focus on sentiment-bearing terms. Tokenisation then splits text into word-level tokens that downstream vectorizers (TF-IDF, Count, Word2Vec, FastText) can consume. Stemming reduces words to their roots (e.g., "membantu" → "bantu"), consolidating morphological variants to improve generalisation. The provided data field text_stemindo reflects these preprocessing steps, so the training pipeline directly vectorises this prepared text.

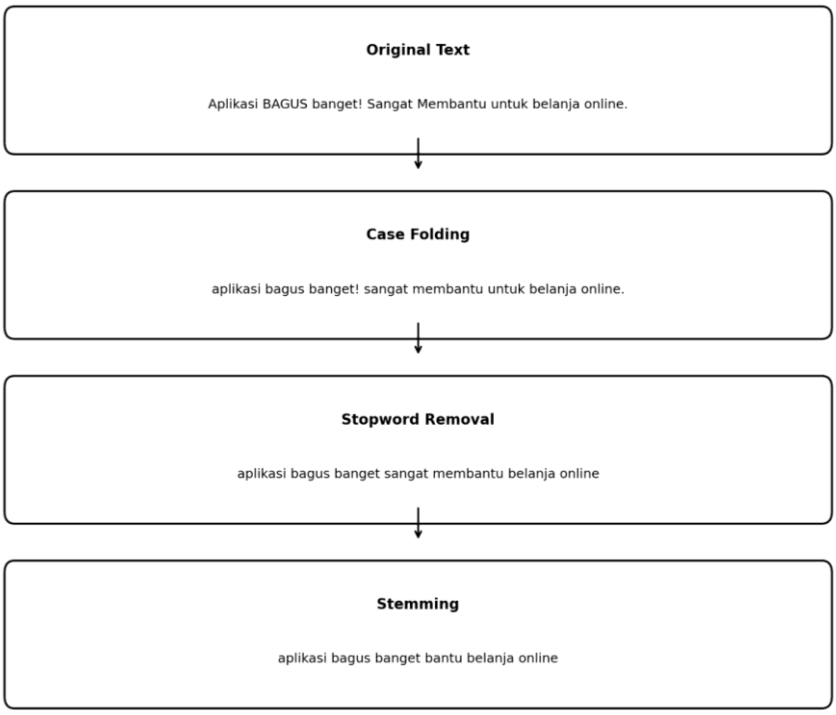

Figure 4. Sequential Text Preprocessing Stages (Case Folding → Stopword Removal → Stemming)

After preprocessing, the cleaned and standardised dataset is partitioned into training and testing subsets for supervised learning and unbiased model evaluation. The dataset contains binary sentiment labels (Positive, Negative). The research adopts an 80:20 split, allocating 80% of the data to model training and reserving 20% for testing. The training set is used to estimate the probabilities required by the Naive Bayes classifier, allowing the model to learn patterns and connections between words and sentiment classes. The test set, which the model hasn't encountered during training, is used to evaluate generalisation performance and to provide an unbiased estimate of classification accuracy. Data splitting is performed using stratified sampling to ensure that the proportion of sentiment classes in both subsets reflects the original distribution in the full dataset. The training set uses the Naïve Bayes classifier to learn sentiment patterns. At the same time, the held-out test provides an unbiased evaluation of generalisation performance.

Subset	Sentiment	Count	Percentage
Training	positif	3647	40.18
Training	negatif	3613	39.81
Testing	positif	891	9.82
Testing	negatif	925	10.19
Total	positif	4538	50.0
Total	negatif	4538	50.0

Figure 5. Distribution of Training and Testing Data by Sentiment Class

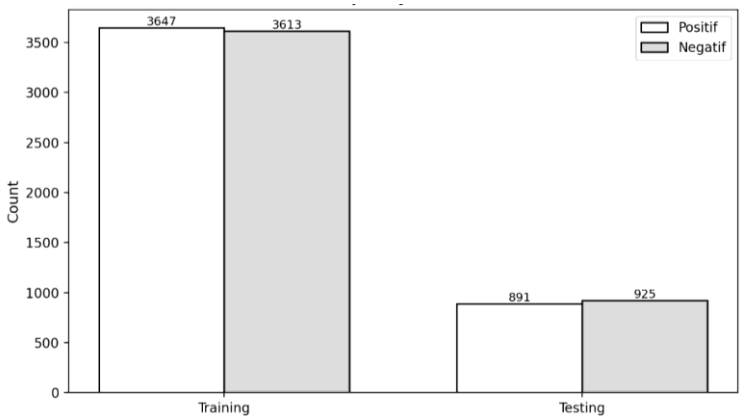

Figure 6. Comparison of Training and Testing Data by Sentiment Class

Feature extraction transforms preprocessed text into numerical representations that machine learning algorithms can process. This research employs Term Frequency-Inverse Document Frequency (TF-IDF) as the feature representation method. TF-IDF assigns weights to each term based on its frequency within a document and its rarity across the entire corpus, thereby highlighting terms that are informative for distinguishing between sentiment classes.

The TF-IDF score for a term *t* in document *d* is calculated as:

$$TF - IDF(t,d) = TF(t,d)\ x\ IDF(t) \quad (1)$$

Where :

- **TF (Term Frequency)** measures how frequently term *t* appears in document *d*:

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \quad (2)$$

- **IDF (Inverse Document Frequency)** measures how rare or common term *t* is across all documents:

$$IDF(t) = \log \frac{\left(Total\ number\ of\ documents\right)}{(Number\ of\ documents\ containing\ term\ t)} \quad (3)$$

This research employs **TF-IDF** with the **TfidfVectorizer** from the scikit-learn library in Python. The vectorizer automatically tokenizes the preprocessed text (from the text_stemindo field), computes TF-IDF scores, and generates a sparse matrix representation where each row corresponds to a review and each column represents a unique term in the vocabulary. This vectorized representation serves as input to the Multinomial Naïve Bayes classification model.

Review	Term 1	Term 2	Term 3	Term 4	Term 5
Review 1	suu (0.707)	banget (0.707)	susah (0.000)	plyletet (0.000)	shopee (0.000)
Review 2	shopee (0.447)	asuh (0.447)	afdet (0.447)	buka (0.447)	mulu (0.447)
Review 3	ok (1.000)	suu (0.000)	susah (0.000)	shopee (0.000)	plyletet (0.000)
Review 4	bagus (1.000)	susah (0.000)	shopee (0.000)	plyletet (0.000)	suu (0.000)
Review 5	susah (0.500)	plyletet (0.500)	knpa (0.500)	aktipkan (0.500)	suu (0.000)

Figure 7. Sample TF-IDF Feature Matrix Showing Top-Weighted Terms

III. RESULTS AND DISCUSSION

The implementation of the sentiment analysis system for Shopee application reviews on the Google Play Store was conducted using the Naive Bayes algorithm. The initial stage involved data acquisition through web scraping techniques, which successfully collected a total of 25,000 user reviews from the year 2024. This dataset size is significantly larger than previous studies, such as Ramadhan et al[6]. (500 reviews) and Rismansyah et al[20]. (2,000 reviews), providing a more robust foundation for training and evaluation. The raw data included attributes such as username, rating score, timestamp, and review content, of which the content and score were the primary focus for analysis. Following data collection, preprocessing was performed to ensure the quality of the textual data. This involved case folding, stopword removal, and stemming, which effectively reduced noise and standardized the vocabulary. The result was a clean dataset where irrelevant characters, numbers, and symbols were removed, leaving only meaningful terms for the model to learn. Table 1 illustrates samples of the data before and after preprocessing, highlighting the transformation from raw, unstructured text to structured tokens ready for vectorization

TABLE I
SAMPLE OF DATA BEFORE AND AFTER PREPROCESSING

Original Review	Preprocessed (Stemmed)	Label
bagus banget	bagus banget	Positive
aplikasi tolol daftar nomr hp aj susah bngt anjg	aplikasi tolol daftar nomr hp aj susah bngt anjg	Negative
cuman di shopee saja langganan belanja saya selalu gratis ongkir 😊 ✌️	cuman shopee langgan belanja gratis ongkir	Positive
sangat berguna dan bermanfaat sebagai sarana jual dan beli	guna manfaat sarana jual beli	Positive
mantap semantap mantapnya. is the best.	mantap mantap mantap is the best	Positive

The labeling process was automated based on the user's rating score, a method chosen to eliminate subjective bias in manual annotation. Reviews with ratings of 1 to 3 were categorized as 'Negative', while ratings of 4 and 5 were classified as 'Positive'. This scoring threshold aligns with the methodology used in similar e-commerce sentiment studies. The resulting class distribution revealed a significant imbalance, characteristic of real-world product reviews: 82.4% of the dataset was labeled 'Positive' (approximately 20,595 reviews), while 17.6% was labeled 'Negative' (approximately 4,405 reviews). To evaluate the model's performance, the dataset was split into training and testing sets using an 80:20 ratio. This split yielded 20,000 reviews for training the Multinomial Naive Bayes model and 5,000 for testing. The split was stratified to ensure that the 82:18 class distribution was preserved in both the training and testing sets, thereby preventing bias arising from an unrepresentative test set.

The feature extraction process utilized the Term Frequency-Inverse Document Frequency (TF-IDF) method. This approach assigned weights to words based on their importance, downweighting standard terms that occur frequently across documents while highlighting unique terms that distinguish optimistic from negative sentiment. The vectorized data were then fed into the Multinomial Naive Bayes classifier, a variant of Naive Bayes explicitly designed for text classification tasks in which features represent word counts or frequencies. The experimental results on the test set of 5,000 reviews demonstrated a high overall performance. The model achieved an Accuracy of 91.96%, correctly classifying 4,598 out of 5,000 reviews. This result indicates that the Naïve Bayes algorithm is highly effective for this domain, even when handling the noisy and informal language commonly found in Indonesian e-commerce reviews.

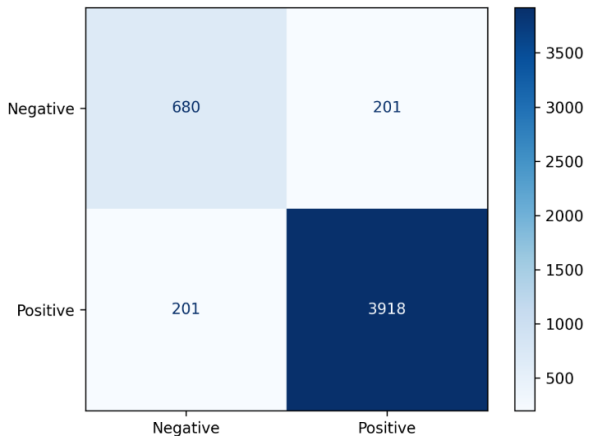

Figure 8. Confusion Matrix Visualization

A deeper analysis of the model's performance was conducted using the Confusion Matrix, as shown in Figure 8. The matrix reveals that the model correctly identified 3,918 True Positives (Positive reviews correctly predicted as Positive) and 680 True Negatives (Negative reviews correctly predicted as Negative). The errors were evenly distributed, with 201 False Positives (Negative reviews predicted as Positive) and 201 False Negatives (Positive reviews predicted as Negative). The balanced nature of the errors (201 FP vs 201 FN) is a significant finding. It suggests that the model does not exhibit a strong bias toward overpredicting one class over the other, despite the pronounced class imbalance in the training data. This balance is crucial for a recommendation system, as it minimizes the risk of suppressing valid complaints (False Positives) or missing out on positive feedback (False Negatives).

The classification report provides further insight into the model's handling of the minority 'Negative' class. The Precision for the Negative class was 77.19%, meaning that when the model predicts a review is negative, it is correct 77.19% of the time. Similarly, the Recall for the Negative class was 77.19%, indicating that the model successfully captured 77.19% of all actual negative reviews in the test set. The F1-Score for the Negative class, which is the harmonic mean of Precision and Recall, also stood at 77.19%. In the context of imbalanced learning, where models often achieve high accuracy by ignoring the minority class, this F1-Score is a strong indicator of the model's robustness. This demonstrates that the high overall accuracy (91.96%) is not merely the result of the model guessing the majority class, but rather stems from a genuine ability to distinguish among sentiment classes.

For the majority of the 'Positive' class, the model performed exceptionally well, as expected. The Precision and Recall were both approximately 95%, with an F1-score of 95%. This high performance on the positive class drives the overall accuracy, but does not overshadow the model's competent handling of the negative class. The weighted average F1-Score across both classes was 92%, reflecting the model's overall reliability. Comparing these results with previous studies highlights the analytical contribution of this research. While Rismansyah et al. [15] achieved an 83% accuracy using Naïve Bayes on 2,000 reviews, our model yielded a near 9% improvement (91.96%). This enhancement is not merely a function of dataset size (25,000 reviews), but rather the model's increased exposure to a broader spectrum of morphological variations, which allows the TF-IDF vectorizer to build a more robust feature space than possible with smaller corpora.

Furthermore, while our 91.96% accuracy aligns with the best-case scenario reported by Ramadhan et al. [6] (92%), their study was fundamentally constrained by a small sample of 500 reviews, making it susceptible to overfitting and high variance in real-world deployments. Achieving this same accuracy ceiling on a scale 50 times larger validates the stability of the Naïve Bayes and TF-IDF pipeline, proving its resilience against the inherent noise and varied syntactic structures typical of massive user-generated content. In contrast, Sitorus and Zufria [1] reported a significantly lower 71% accuracy using a 3-class structure (Positive, Negative, Neutral). Our binary approach explicitly eliminates the syntactic overlap and ambiguous decision boundaries introduced by neutral user sentiment. Because natural language in e-commerce tends to be strongly polarized, simplifying the decision boundary allowed our Multinomial Naïve Bayes model to distinctly separate the feature probabilities of complaints versus praises, thereby maximizing both overall accuracy and minority class capture.

The visualization of the results through Bigram Analysis (Figure 9 and Figure 10) further supports the quantitative findings by highlighting the most frequent two-word combinations. Unlike single-word frequencies, Bigrams provide context, revealing specific aspects of user satisfaction or dissatisfaction. The Positive Bigram chart (Figure 9) is dominated by phrases such as "pengiriman cepat" and "barang bagus", reflecting user satisfaction with delivery speed and product quality. Conversely, the Negative Bigram chart (Figure 10) highlights phrases like "pengiriman lama" and "tidak sesuai", pointing to specific pain points regarding shipping delays and product discrepancies.

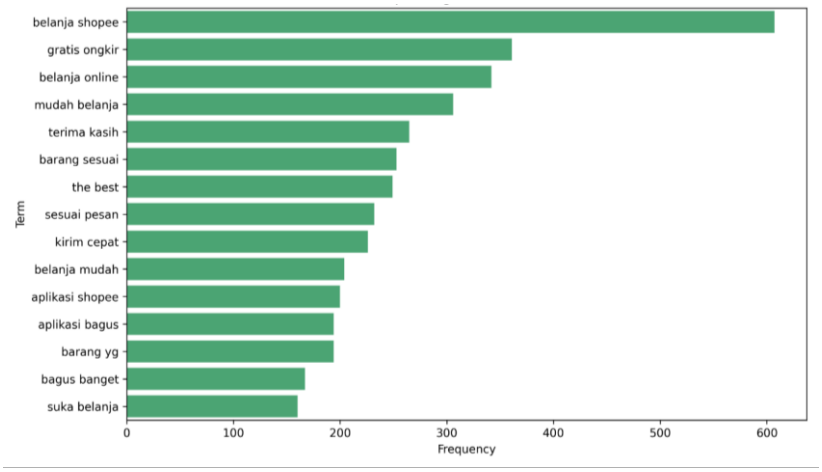

Figure 9. Top 15 Bigrams Positive Sentiment

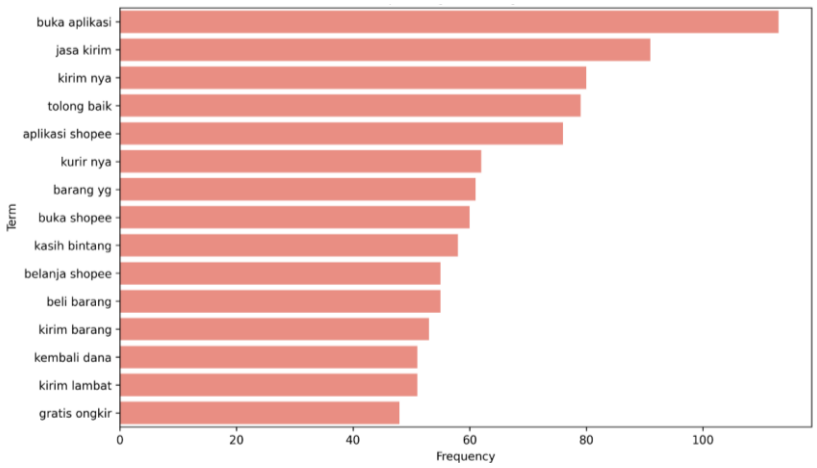

Figure 10. Top 15 Bigrams Negative Sentiment

Despite the strong overall accuracy, a critical limitation of this study lies in the 77% recall rate for the negative class. While substantial for highly imbalanced data (82% positive vs. 18% negative), this indicates that 23% of actual user complaints are still misclassified as positive. This shortfall is largely attributed to the inherent noise in user-generated text, such as sarcasm (e.g., giving a high rating but writing "Bagus banget barangnya, baru dipakai sehari langsung rusak"), ambiguous slang, and the simplicity of the model. Furthermore, the automated rating-based labeling method, while highly scalable, occasionally misaligns with the nuanced textual sentiment. Consequently, while Multinomial Naïve Bayes provides a robust baseline, it inherently struggles to perfectly resolve the severe class imbalance without the aid of advanced resampling techniques or more complex deep learning architectures.

Nevertheless, these findings yield significant practical implications for e-commerce business intelligence. By successfully identifying 77% of negative feedback within a massive, imbalanced dataset, the model acts as an automated, large-scale filter to isolate critical pain points—such as shipping delays ("pengiriman lama")—enabling developers and customer service teams to prioritize interventions efficiently. Ultimately, the model's performance confirms that Naïve Bayes is an optimal, lightweight solution for initial data imbalance handling. It provides a reliable and highly scalable framework for real-time decision-making in e-commerce, balancing computational efficiency with actionable analytical depth.

IV. CONCLUSION

This study successfully implemented the Multinomial Naïve Bayes algorithm to analyze sentiment across 25,000 Shopee application reviews, significantly expanding upon previous small-scale research. The model achieved a robust classification accuracy of 91.96%, validating the scalability of Naïve Bayes for massive, noisy Indonesian text datasets. Crucially, the research effectively handled severe real-world class imbalance by maintaining a strong 77% recall rate for the negative minority class, ensuring critical user complaints are captured for practical business intelligence. Bigram analysis further enriched these findings by identifying specific textual drivers of user satisfaction and dissatisfaction, notably highlighting logistics and delivery speed as primary concerns. To build upon this foundation, future work should explore Aspect-Based Sentiment Analysis (ABSA) to automatically categorize feedback into functional dimensions such as 'Logistics', 'Product Quality', or 'App Usability'. Furthermore, comparative studies utilizing advanced deep learning architectures, such as LSTM or BERT, should be pursued to better resolve the complexities of sarcastic, ambiguous, and slang-heavy e-commerce reviews.

SPECIAL THANKS

The author would like to express the deepest gratitude to all parties who assisted in the completion of this research. Special appreciation is extended to the affiliated institution, academic supervisors, and relevant technical support teams for the facilities, developmental resources, and constructive guidance provided throughout the duration of this study. Their continuous institutional and academic support has been instrumental in the successful implementation and finalization of this research.

DAFTAR PUSTAKA

[1] R. A. Sitorus and I. Zufria, "Application of the Naïve Bayes Algorithm in Sentiment Analysis of Using the Shopee Application on the Play Store," *Digit. Zone J. Teknol. Inf. Dan Komun.*, vol. 15, no. 1, pp. 53–66, May 2024, doi: 10.31849/digitalzone.v15i1.19828.

[2] N. A. Maulana and Z. Fatah, "Penerapan Metode Naïve Bayes untuk Analisis Sentimen Ulasan Produk di Platform E-Commerce," Global Journal of Management and Informatics (GJMI), vol. 2, no. 11, pp. 433-439, 2023.

[3] "T. D. Darmawan, "Analisis Sentimen Ulasan Pengguna Aplikasi Shopee Menggunakan Metode Multinomial Naïve Bayes," Undergraduate Thesis, Universitas Dinamika, Surabaya, 2022.

[4] R. D. Kurniawan, A. Yohannis, and W. T. Atmojo, "Sentiment Analysis of Getcontact Application Reviews on Google Play Store Using Naive Bayes Algorithm," *J. Tek. Inform. Jutif*, vol. 6, no. 4, pp. 2848–2858, Sep. 2025, doi: 10.52436/1.jutif.2025.6.4.5248.

[5] S. Fide, S. Suparti, and S. Sudarno, "ANALISIS SENTIMEN ULASAN APLIKASI TIKTOK DI GOOGLE PLAY MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) DAN ASOSIASI," *J. Gaussian*, vol. 10, no. 3, pp. 346–358, Dec. 2021, doi: 10.14710/j.gauss.v10i3.32786.

[6] B. Z. Ramadhan, R. I. Adam, and I. Maulana, "Analisis Sentimen Ulasan pada Aplikasi E-Commerce dengan Menggunakan Algoritma Naïve Bayes," *J. Appl. Inform. Comput.*, vol. 6, no. 2, pp. 220–225, Dec. 2022, doi: 10.30871/jaic.v6i2.4725.

[7] R. Kurniawan, H. O. L. Wijaya, and R. P. Aprisusanti, "Sentiment Analysis of Google Play Store User Reviews on Digital Population Identity App Using K-Nearest Neighbors," *J. Sisfokom Sist. Inf. Dan Komput.*, vol. 13, no. 2, pp. 170–178, Jun. 2024, doi: 10.32736/sisfokom.v13i2.2071.

[8] C. Apriansyah Hutagalung and V. Budi Lestari, "Data Mining Approach: K-Means Clustering and Naïve Bayes Classifier for Graduate Quality Analysis," *J-KOMA J. Ilmu Komput. Dan Apl.*, vol. 8, no. 1, pp. 33–42, Jun. 2025, doi: 10.21009/j-koma.v8i1.05.

[9] 9] M. R. Firdaus, N. Rahaningsih, and R. D. Dana, "Analisis Sentimen Aplikasi Shopee di Google Play Store Menggunakan Klasifikasi Algoritma Naïve Bayes," Jurnal Informatika dan Rekayasa Perangkat Lunak, vol. 6, no. 1, 2024.

[10] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory," Feb. 14, 2017, *arXiv*: arXiv:1410.5329. doi: 10.48550/arXiv.1410.5329.

[11] Vidhya. K. A and G. Aghila, "A Survey of Naïve Bayes Machine Learning approach in Text Document Classification," 2010, *arXiv*. doi: 10.48550/ARXIV.1003.1795.

[12] I. H. Kusuma and N. Cahyono, "Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor," *J. Inform. J. Pengemb. IT*, vol. 8, no. 3, pp. 302–307, Sep. 2023, doi: 10.30591/jpit.v8i3.5734.

[13] R. A. Prasetyo, "Perbandingan Algoritma Logistic Regression, SVM, dan Random Forest pada Analisis Sentimen Aplikasi Gopay," *J. Inform. J. Pengemb. IT*, vol. 10, no. 4, pp. 1176–1188, Sep. 2025, doi: 10.30591/jpit.v10i4.8796.

[14] T. Salsabilla and D. Alita, "Analisis Sentimen Inses di Social Media menggunakan Algoritma Naïve Bayes," *J. Inform. J. Pengemb. IT*, vol. 9, no. 3, pp. 271–280, Dec. 2024, doi: 10.30591/jpit.v9i3.6611.

[15] I. Septiana and D. Alita, "Perbandingan Random Forest dan SVM dalam Analisis Sentimen Quick Count Pemilu 2024," *J. Inform. J. Pengemb. IT*, vol. 9, no. 3, pp. 224–233, Dec. 2024, doi: 10.30591/jpit.v9i3.6640.

[16] A. Adiuntoro and A. Hendrawan, "Klasifikasi Pertanyaan Quora Menggunakan Metode Keyword-based dan Analisis Sentimen dengan ComplementNB," *J. Inform. J. Pengemb. IT*, vol. 10, no. 2, pp. 432–439, Apr. 2025, doi: 10.30591/jpit.v10i2.7965.

[17] *Google Play Store*. (Dec. 12, 2025). [Online]. Available: https://play.google.com/store

[18] A. Gondowastradmojo and R. Kusumastuti, "Analisa Sentimen Ulasan di Tokopedia dengan Metode Naïve Bayes," SEMNASA, 2024.

[19] S. Dermawan and A. T. Ayunda, "Sentiment Analysis of Coretax on Social Media X Using Naive Bayes, SVM, and LSTM for Service Improvement," Journal of Applied Informatics and Computing (JAIC), vol. 9, no. 6, 2025..

[20] R. R. Rismansyah, A. Sudiarjo, and T. Mufizar, "ANALISIS SENTIMEN ULASAN SHOPEE PADA GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA NAIVE BAYES," *JEIS J. Elektro Dan Inform. Swadharma*, vol. 5, no. 1, pp. 109–120, Jan. 2025, doi: 10.56486/jeis.vol5no1.661.