JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 3, Juni 2025, hlm. 457 – 462 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i3.3637
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

SENTIMENT ANALYSIS USING NAIVE BAYES ALGORITHM CASE STUDY ON AMAZON E-COMMERCE PRODUCT REVIEWS

Erik Rahman[1*], Namora[2], Lukman Anas[3]
[1]Bisnis dan Teknologi, Universitas Pertiwi
[2]Bisnis dan Teknologi, Universitas Pertiwi
[3]Teknik, Universitas Muhammadiyah Makassar
email: [*]erik.rahman@pertiwi.ac.id

Abstract: Sentiment analysis techniques involve the process of recognizing and categorizing viewpoints in text into groups such as positive, negative, or neutral. This research aims to evaluate product reviews on the Amazon e-commerce platform using the Naive Bayes algorithm. The product review dataset used was obtained from Kaggle, consisting of 5,000 product reviews labeled into three sentiment categories: positive, negative, and neutral. The methodology includes stages of text pre-processing (cleaning, case folding, tokenization, stopword removal, and stemming), feature extraction using TF-IDF technique, and classification using the Naive Bayes algorithm. Model evaluation is conducted using metrics of accuracy, precision, recall, and F1-score. The results show an accuracy of 94%. These findings indicate that the Naive Bayes algorithm is effective in classifying the sentiment of product reviews. The results of this research can help business actors in automatically analyzing consumer opinions as a basis for decision-making.

Keywords: e-commrece; naive bayes; product reviews; tf-idf; sentiment analysis.

Abstrak: Teknik analisis sentimen merupakan proses mengenali dan mengkategorikan sudut pandang dalam teks menjadi kelompok-kelompok seperti positif, negatif, atau netral. Penelitian ini bertujuan untuk menilai ulasan produk pada platform e-commerce Amazon dengan menggunakan algoritma Naive Bayes. Dataset ulasan produk yang digunakan diperoleh dari Kaggle, terdiri dari 5.000 ulasan produk yang telah dilabeli ke dalam tiga kategori sentimen: positif, negatif, dan netral. Metodologi meliputi tahapan pra-pemrosesan teks (cleaning, case folding, tokenisasi, stopword removal, dan stemming), ekstraksi fitur menggunakan teknik TF-IDF, serta klasifikasi menggunakan algoritma Naive Bayes. Evaluasi model dilakukan dengan menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil menunjukkan akurasi mencapai 94%. Temuan ini menunjukkan bahwa algoritma Naive Bayes efektif dalam mengklasifikasikan sentimen ulasan produk. Hasil dari penelitian ini dapat membantu pelaku bisnis dalam menganalisis opini konsumen secara otomatis sebagai dasar pengambilan keputusan.

Kata kunci: analisis sentimen; e-commerce; naive bayes; tf-idf; ulasan produk.

INTRODUCTION

In the digital era, user reviews play a vital role in shaping customer perceptions of products. With the large volume of review data, manual analysis becomes inefficient. Therefore, the analy sis process is automated using the Naive Bayes machine learning algorithm. The use of Naive Bayes aims to categorize product reviews on e-commerce sites into positive and negative reviews is the main emphasis of this study.

This study uses pre-processing stages including data cleaning, case folding, stopword removal, tokenization, and stemming. Feature ex-traction is carried out using the TF-IDF method, and model

DOI: http://dx.doi.org/10.33330/jurteksi.v11i3.3637
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

evaluation uses accu-racy, precision, re-call, and F1-score met-rics. This stage allows the system to un-derstand the con-text of words numerical-ly and improve classification perfor-mance. This study is supported by vario us previous studies, including Bird et al. (2009), Pang & Lee (2008), Jurafsky & Martin (2023), and Salton & McGill (1986). Similar studies were also con-ducted by Aggarwal & Zhai (2012) in text mining and Mikolov et al. (2013) on vector-based word repre-sentation, demon strated the relevance and effec-tiveness of this approach in various NLP domains.

Previous studies have shown that Naive Bayes is often used for text classi-fication due to its simplicity and efficien-cy [1 Opinion mining, also kno wn as sentiment analysis, is a field of study dedicated to examining the thoug hts, impressions, and opinions of indivi duals about various topics, products, or ser-vices. This analysis primarily seeks to automate the identification of sentiments and opinions expressed by people, and then categorize the polarity of these sen-timen[1]. Since the early 2000s, this field has emerged as one of the most dynamic research domains in Natural Language Processing (NLP)[2]. The pace of activi-ty in the field of sentiment analysis has skyrocketed, with its applications spread-ing across a wide range of areas includ-ing business, healthcare, elections, prod-uct evaluation, and market research. Re-views serve as evaluations given by indi-viduals regard-ing a particular product or service.

By leveraging sentiment analysis, customer perspectives can be extracted from these reviews, transforming them into mean-ingful insights for potential buyers. For example, a large number of individuals often read reviews before making a pur-chasing decision. Accor ding to a survey by BrightLocal, 79% of consumers re-ported that they trust online reviews more than personal endorsem ents from rela-tives or friends[3]. Various techniques used in sentiment analysis in-clude machine learning techniques, lexi-con-based approaches, hybrid models, and addition-al strategies such as aspect-based meth-ods, transfer learning, and multimodal sentiment analysis[4].

The increasing amount of user-generated content has triggered the need for automatic sentiment analysis. With the emergence of deep learning and pre-trained models like BERT, sentiment classification has achieved higher accura cy compared to traditional models[5], [6]. This study applies sentiment classifica-tion through machine learning techniques that utilize the Naive Bayes algorithm,by utilizing user review data from the Ama-zon e-commerce web-site available on Kaggle. The choice of machine learning methods is motivated by their efficiency in data analysis, which significantly re-duces the time and effort required to clas-sify reviews. Evaluating the performance of the Naive Bayes algo-rithm in senti-ment categorization is the goal of this study.

METHOD

The research framework consists of several main stages:[1] Data Colle ction: Publicly available product review datasets were used as the basis for the research. The dataset, derived from the Amazon product reviews collection on Kaggle, contains structured sentiment lab els that align with previous benchm arks in sentiment classification[1]; [2]. Prepro cessing Stage: Perform cleaning and pro-cessing of text data to ensure the quality of model input.

We employed NLTK and SpaCy for preprocessing, which are widely used libraries for text normalization and toke

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 3, Juni 2025, hlm. 457 – 462 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i3.3637
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

nization tasks in NLP pipelines[7]; [3] Extraction Feature: Converts text into numeric representation using TF-IDF method; [4]Application of Naive Bayes Model: Using probabilistic models for text classification; [8]Model Evaluation: Using criteria such as accuracy, precisi on, recall, and F1-score to measure mod el performance; [9]Results Analysis: Comparing model performance with other algorithmic approaches.

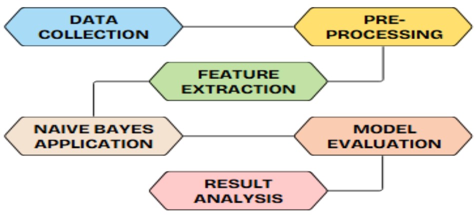
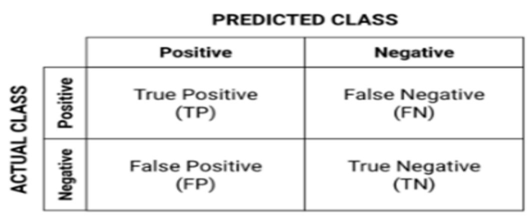


Image 1. Research Framework

Sentiment analysis, often referred to as opinion mining, is an area that explores people's views, feelings, evaluations of views, feelings, and attitudes about things, such as goods, services, organiza-tions, people, events, issues, or subjects discussed in writing. This field has made significant contributions to research in NLP by introducing previously under-studied challenges and problems. Senti-ment analysis is a subset of semantic analysis, its approach is more targeted and specific. Sentiment analysis only needs to understand certain components, such as other people's views and the items being evaluated, rather than having to 'understand' every line or document.

Image 2. Sentiment Analysis

According to Bird et al. (2009), text preprocessing involves steps such as tokenization, stopword removal, and stemming to simplify the text without losing important information [1]. According to Salton and McGill (1986), TF-IDF is a numerical method for evalu-ating how important a particular word is in a particular document compared to the corpus as a whole [2]. According to Rennie et al. (2003), Naive Bayes is a probabilistic algorithm based on Bayes' Theorem, assuming that fea-tures are independent. This algorithm has proven effective in text analysis using word frequency features. An *n x n* matrix that compares the pre-dicted and actual classifications, where n is the number of classes is called a Confusion Matrix. This technique is used to summarize the performance of the algorithm based on classification. As shown in Image 3, a number of terms from the confusion matrix, including: (TP) True Positive, shows the number of correct negative predictions; (FP) False Positive, shows the number of incorrect positive predictions; (TN) True Negative, shows the number of incorrect negative predictions; (FN) False Negative, shows the number of correct positive predictions.



JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XI No 3, Juni 2025, hlm. 457 – 462 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i3.3637
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Image 3. Confusion Matrix Results [7]

Accuracy, precision, recall, and F1-score can be calculated by utilizing data from the confusion matrix. Accuracy serves to assess the extent to which the model successfully classifies data accurately, measured by dividing the total accurate predictions by the total samples. On the other hand, precision indicates the degree of agreement between the desired data and the model's prediction results; a high precision value reflects a small number of false positives. F1-score, which is a weighted harmonic mean of precision and recall, is a statistic created by combining the two variables. Together with its calculation method, Image 4 offers a measure to evaluate the effectiveness of a machine learning model.

Metrik	Formula
Akurasi	$\frac{TP + TN}{TP + TN + FP + FN}$
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
F1-score	$2 * \frac{Recall * Precision}{Recall + Precision}$

Image 4. Classification Evaluation Metrics[10]

RESULTS AND DISCUSSION

Preprocessing Stage

The preprocessing process is done using Python libraries such as NLTK and SpaCy: a) Data Cleaning: Removes punctuation, numbers, and special characters[1]. Example: Before: "This product is good!!! But the packaging is bad :("; After: "This product is good but the packaging is bad"; b) Text Normalization: Changes capital letters to lower case; c) Tokenization: Splitting text into words. d) Stopword Removal: Removes common terms such as "and", "in", "the" [1]; e) Stemming: Changing words to their basic words to maintain uniformity [1]. Research design, but focused on the results of research that has been carried out.

Extraction Features

As shown in Image 2, the review text is represented as a numeric vector using the Term Frequency-Inverse Document Frequency (TF-IDF) approach. In TF-IDF, the frequency of occurrence of a word in a given document (TF) is calculated and adjusted based on the frequency of occurrence of the term in all documents (IDF) [2]. Although neural embeddings like Word2Vec and BERT have gained popularity for capturing contextual semantics, TF-IDF remains a robust baseline for text representation in short-review datasets. Its simplicity, interpretability, and effectiveness in high-dimensional sparse text classification tasks make it a suitable choice for this study [10], [11]. Moreover, TF-IDF requires less computational resources and training time, making it ideal for scenarios where quick implementation and reliable performance are prioritized over deep contextual understanding.

Tabel 1. TF-IDF

Say	TF-IDF Value
Product	0.45
Good	0.34
Packaging	0.21
Bad	0.19

Naïve Bayes Algorithm

The Naive Bayes algorithm works based on Bayes Theorem, where: Class probability given data; Data probability in class; Initial probability of class; Data probability. Naive Bayes assumes that all features are independent, thus allowing simple yet effective probability calculations for word classification [3]. The implementation of sentiment classi-fication using the Naive Bayes method begins with the model creation process, which includes importing modules and classes available in the sklearn library. BernoulliNB, ComplementNB, GaussianNB, and MultinomialNB are some of the types of Naive Bayes classifiers used. The next stage is modeling, which involves feeding training data to the model after the model to be used has been selected. Using test data to predict sentiment labels is the final step.

Model Evaluation

80% of the dataset is used for training, while 20% is used for testing. K-fold cross-validation (k=5) is used for validation. The evaluation metrics used include: Accuracy: Percentage of accurate predictions compared to total test data; Precision: Ratio of accurate positive predictions to all positive predictions. Recall: Ratio of actual positive data to the number of accurate positive predictions. F1-Score: Harmonic mean value between precision and recall. The results of model evaluation on test data based on the confusion matrix are as follows:

Tabel 2. Confusion Matrix Results

	Positive Prediction	Negative Prediction
Positive Current	950	50
Negative Current	70	930

The Model Evaluation Matrix is described in Table 3.

Tabel 3. Model Evaluation Matrix

	Percentage
Accuracy	94
Precision	93
Recall	95
F1-Score	94

Compared to Logistic Regression and SVM, Naive Bayes remains competitive due to its simplicity and low computational cost [12], [13].

Results Analysis

Naive Bayes showed excellent per-formance with 94% accuracy, 93% precision, and 95% recall. This shows that the model has a low error rate when classifying reviews, both posi-tive and negative. However, there are some limitations: 1. Minor Classification Error: A total of 70 negative reviews were misclassified as positive. This could be due to the use of ambiguous words in the reviews; 2. Ambiguous Reviews: Some reviews are neutral or contain ambiguous context (e.g., “good prod-uct, but slow delivery”) and are difficult for the model to classify well. For comparison, Logistic Regres-sion and Support Vector Machine (SVM) algorithms are also applied to the same dataset. The accuracy results of each algorithm are shown in the following table 4.

Tabel 4. Algorithm Accuracy Results

Alghoritm	Accuracy
Naive Bayes	94%
Logistic Regression	95%
SVM	96%

From the confusion matrix table, it can be explained that: (TP) True Positives: 950 (positive evaluation results categorized as positive); (TN) True Negatives: 930 (negative evaluation results categorized as negative); (FP) False Positives: 70 (negative evaluation results incorrectly classified as positive); (FN) False Negatives: 50 (positive evaluation results incorrectly categorized as negative). Future integration with attention-based architectures such as BiLSTM and CNN may improve the model's ability to capture contextual dependencies in sentiment-laden texts [14], [15].

CONCLUSION

The results of this study demonstrate the reliable capacity of the Naive Bayes algorithm to perform sentiment analysis on Amazon e-commerce product evaluations. The model achieved 94%

accuracy, 93% precision, 95% recall, and 94% F1 score with pre-processing stages including data cleaning, de-capitalization, tokenization, common word removal, and stemming, as well as feature extraction using the TF-IDF approach. The evaluation metrics showed that the model performed well in classifying and had a low error rate. Although the performance of this model was very good, the study findings also showed that this model has limitations in handling ambiguous or neutral reviews. Therefore, it is recommended that future studies investigate embedding-based word representation methods such as Word2Vec or BERT.

These embedding methods can better understand the meaning of words in various contexts by capturing the deeper semantic context of words in sentences. It is anticipated that the application of this technique will improve the classification accuracy, especially in reviews containing complex linguistic patterns and multiple meanings.

BIBLIOGRAPHY

[1] UCI Machine Learning Repository, "Sentiment Analysis Dataset," 2024.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson, 2023.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[4] A. Sharma, S. Kaur, N. Memon, A. J. Fathima, S. Ray, and M. W. Bhatt, "Alzheimer's patient detection using support vector machine (SVM) with quantitative analysis," *Neurosci. Informatics*, vol. 1, no. 3, p. 100012, 2021.

[5] Y. Zhou, Q. Wu, and Y. He, "A hybrid deep learning model for sentiment analysis based on BERT and CNN," *Expert Syst. Appl.*, vol. 167, p. 114103, 2021.

[6] M. Cambria, D. Das, S. Bandyo padhyay, and A. Feraco, *A Practical Guide to Sentiment Analysis*. Springer, 2022.

[7] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 12, no. 2, pp. 1–183, 2020.

[8] J. Meiyazhagan, S. Sudharsan, A. Venkatasen, and M. Senthilvelan, "Prediction of occurrence of extreme events using machine learning," *Procedia Comput. Sci.*, vol. 192, pp. 548–557, 2021.

[9] A. Al-Smadi, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," *J. Inf. Knowl. Manag. Syst.*, vol. 49, no. 1, pp. 17–33, 2021.

[10] S. Dey, A. S. Rath, and S. Mohanty, "An optimized TF-IDF based feature extraction strategy for sentiment classification using deep learning," *J. King Saud Univ. - Comput. Inf. Sci.*, 2023.

[11] R. M. Alguliyev, R. M. Aliguliyev, and L. A. Abdullayeva, "A review of sentiment analysis research in the era of deep learning," *Expert Syst. Appl.*, vol. 169, p. 114376, 2021.

[12] A. Rani, S. S. Arora, and V. Bhatnagar, "Comparative evaluation of machine learning algorithms for sentiment analysis," *Procedia Comput. Sci.*, vol. 184, pp. 330–338, 2021.