The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 4, August 2025
https://doi.org/10.33022/ijcs.v14i4.4980

Machine Learning and Transformer-based Model for Sentiment Analysis of Indonesian E-Commerce Reviews

Wahyu Widyananda[1*], Maskur[2], Ahmad Fauzi[3]
wahyu.widyananda@polinema.ac.id[1], maskur@polinema.ac.id[2], ahmad.fauzi@polinema.ac.id[3]
[1,2,3]State Polytechnic of Malang, Malang, Indonesia

Article Information

Received : 21 Aug 2025
Revised : 27 Aug 2025
Accepted : 30 Aug 2025

Keywords

Machine Learning, Transformer-based Model, Sentiment Analysis, E-commerce, Indonesian Language

Abstract

The growth of e-commerce in Indonesia has produced a large volume of user-generated reviews, which contain valuable knowledge for business decisions. However, analyzing this unstructured text data manually is inefficient. The purpose of this study is to improve the performance of sentiment classification on Indonesian e-commerce reviews using machine learning and transformer-based models. The test method is carried out using a public e-commerce review dataset. Three models are evaluated: Multinomial Naïve Bayes, Support Vector Machine (SVM), and IndoBERT. For machine learning models, text pre-processing is performed, and features are extracted using TF-IDF. For the transformer-based model, a fine-tuning approach is used. The results show that the IndoBERT model produces better classification accuracy than the other tested models. For the given dataset, this method obtains 94,1% in accuracy, outperforming both SVM (89,5%) and Multinomial Naïve Bayes (84,2%). The IndoBERT model, despite its higher computational cost, is the most effective for this classification task.

Attribution-ShareAlike 4.0 International License

A. Introduction

As the digital economy in Indonesia expands, e-commerce platforms have become a primary source of consumer data [1], [2]. Platforms such Tokopedia and Shopee generate a large amount of user reviews for products and services [3]. The high volume and velocity of this user-generated text often written in informal Indonesian language, present a significant data processing challenge. This data contains consumer opinions that can be used to create predictive models for business intelligence. The effective analysis of these opinions allows companies to track brand perception, identify key product features, and understand market trends. By applying classification methods on this sentiment analysis, business can transform unstructured textual data into meaningful insights that support strategic decision making and enhance customer engagement.

To address this classification challenge, researchers have widely applied supervised machine learning techniques. For example, research on Indonesian Twitter data has shown that the Support Vector Machine (SVM) is highly effective method for sentiment classification [4]. Similarly, studies on Gojek and Grab application reviews have demonstrated that probabilistic classifiers like Naïve Bayes can also achieve strong performance, making them a common baseline for this type of classification [5]. However, although these machine learning algorithms have proven effective, they often depend on bag-of-words feature representation like TF-IDF that ignore the meaning and context of words in a sentence [6]. This can lead to significant errors in prediction, particularly when handling the complexities of informal language.

Recent development in deep learning and transformer architectures like BERT (Bidirectional Encoder Representations) has introduced models that can better understand language context [7]. By using self-attention mechanisms and being pre-trained on vast amount of text, these models can generate contextualized word embeddings that account for semantic and ambiguity. For the Indonesian language, the IndoBERT model has been specificially pre-trained on large local corpus, showing strong performance on various NLP tasks [8]. This language-specific pre-training is critical, as it allows the model to learn the unique grammatical structures and colloquialisms of Indonesian language more effectively than a general multilingual model.

Although there have been studies related to sentiment analysis in Indonesian, a direct performance evaluation of machine learning models against language-specific transformer model on e-commerce data is needed. In fact, selecting an inappropriate model can lead to poor predictive accuracy and wrong business decisions. The IndoBERT approach can be used as a classification method to improve the prediction accuracy of e-commerce review sentiment.

In contrast to previous studies, this study aims to improve the classification performance of Indonesian e-commerce reviews sentiment analysis using a fine-tuned IndoBERT model. This study evaluates its performance against other classification algorithms, such as Multinomial Naïve Bayes and SVM. Parameters used to measure classification performance include accuracy, precision, recall, and F1-score. This research was conducted to overcome the problem of text classification to make business decisions accurately.

For many years, the primary approach for sentiment analysis has been supervised machine learning, which relies on feature engineering from text. The "bag-of-words" model, often weighted using Term Frequency-Inverse Document Frequency (TF-IDF), is a standard method for converting text into numerical format [9]. Among the most widely used classifiers, Multinomial Naïve Bayes is noted for its computational efficiency and strong performance as a baseline model, despite its simplistic assumption that features (words) are conditionally independent [10]. Support Vector Machine (SVM) is another powerful algorithm that has consistently demonstrated high performance in text classification. By mapping features into a high-dimensional space, SVM can find an optimal separating hyperplane, making it highly effective for sparse, high-dimensional data typical of text [11]. Numerous studies in various language have validated the effectiveness of SVM for sentiment analysis [12].

Transformer-based models have changed Natural Language Processing (NLP) by using self-attention to focus on important words in a sentence, helping the model capture meaning across longer text [7]. Building on this, Google's BERT model established a new state-of-the-art for a wide range of NLP tasks. Unlike other machine learning models that see text as a flat bag of words, BERT is pre-trained on a massive text corpus to understand language context bidirectionally. This pre-trained model can then be "fine-tuned" on a smaller, task-specific dataset to achieve very high performance with less labeled data [13].

Recognizing that general multilingual models may not capture the specific nuances of a single language, researchers have developed monolingual BERT models. IndoBERT is a version of the BERT model that has been specifically pre-trained on a massive (over 220 million words) Indonesian dataset called Indo4B [8]. This specialized pre-training allows IndoBERT to have a deeper understanding of Indonesian grammar, slang, and context compared to its multiligual counterparts. Studies has shown that fine-tuning IndoBERT yields superior performance on various Indonesian NLP tasks, including sentiment analysis, text summarization, and named entity recognition [14], [15]. This makes IndoBERT particularly suitable for tasks involving Indonesian text, where understanding local language patterns is essential.

B. Research Method

This study followed a clear, step-by-step process to provide a consistent foundation for evaluating the selected models. The research methodology design can be shown in Figure 1.

Dataset

The dataset used in this study is the "Indonesian Marketplace Product Reviews" dataset [16]. This dataset is publicly available on Kaggle platform. It contains 831 reviews written in Indonesian language, and each labelled with a sentiment classification ('positive' or 'negative'). Specifically, the dataset includes 446 negative reviews and 385 positive reviews.

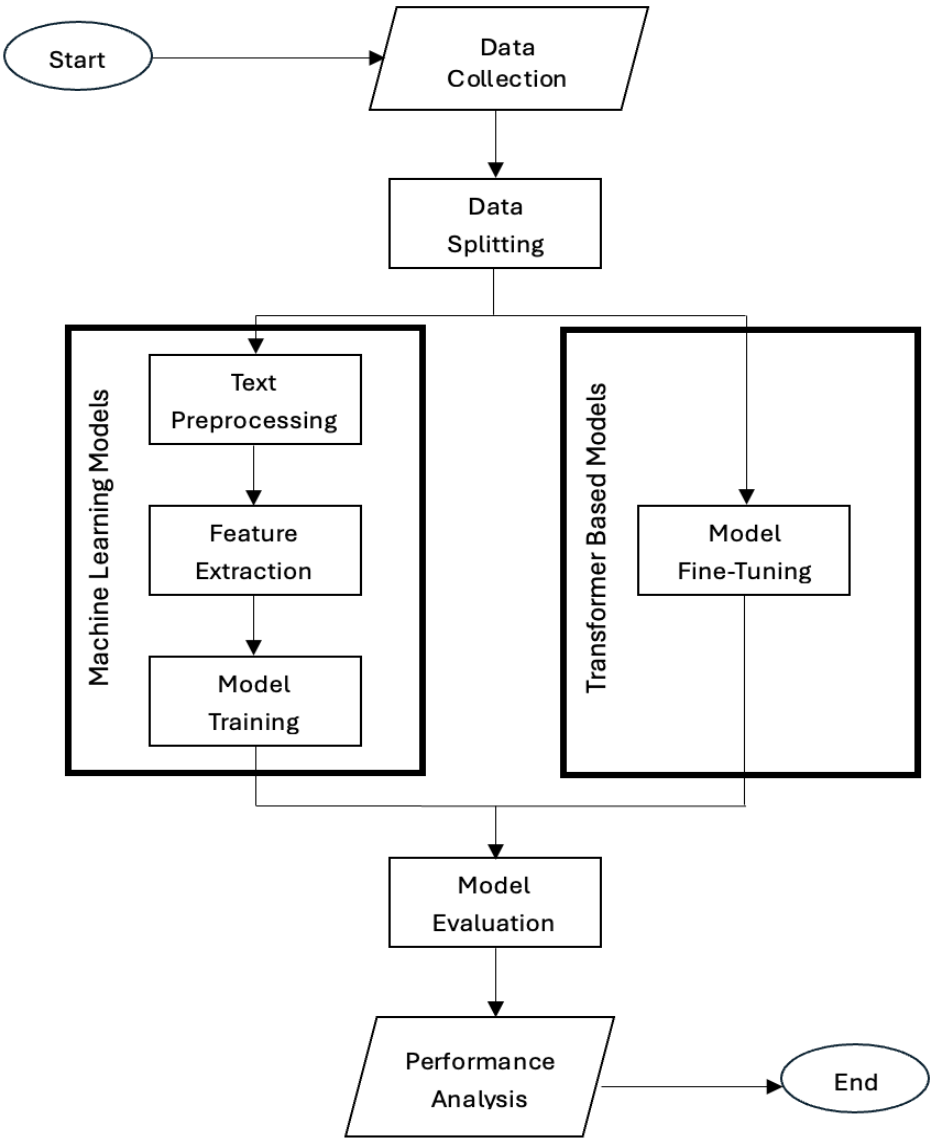

Figure 1. Research Methodology Design

Data Preprocessing

Raw textual data is inherently unstructured and often includes irrelevant or noisy elements. To prepare it for effective use in machine learning models, a set of established NLP preprocessing techniques was employed [17].

1. Case Folding: All text was converted to lowercase to ensure uniformity (e.g. “Andi” and “andi” are treated as the same word)
2. Tokenization: The text was broken down into individual words or token
3. Stopwords Removal: Common Indonesian words that do not carry significant meaning (e.g. “di”, “yang”, “dan”) were removed using a standard Indonesian stopword list. The importance of language-specific stopword list is crucial for effective feature selection [18].
4. Stemming: Words were reduced into their root form (e.g. “membaca” becomes “baca). For this purpose, a popular rule-based stemmer for the Indonesian language called Sastrawi was used [19].

Feature Engineering

A key distinction in this study is the feature engineering approach for the different model types:

- For Multinomial Naïve Bayes and SVM: The pre-processed text was converted into numerical vectors using TF-IDF algorithm. This creates a matrix where each row represents a review and each column represents a unique word in this corpus, with the cell value being the TF-IDF weight [9]. The formula used to compute TF-IDF is presented as follows.

$$W_{t,d} = tf(w,d)\ x\ idf(w,D)$$

  In this formula, $W_{t,d}$ represent the TF-IDF weight of the word w in document d. The term w refers to the number of documents that contain the word, while D denotes the total number of documents in the dataset.

- For IndoBERT: The raw review texts were directly input into the pre-trained IndoBERT model without extensive pre-processing, as the model itself includes its own tokenization mechanism. The model converts the text into dense, contextualized word embeddings. A simple feed-forward neural network was added as classification layer on top of IndoBERT, which is then fine-tuned on sentiment analysis task.

Model Training and Evaluation

The provided dataset was split into 80% for training and 20% for testing to ensure a fair assessment of the model's performance on previously unseen data.

- The Multinomial Naïve Bayes and SVM models were trained on the TF-IDF vectors of the raining set
- The IndoBERT model was fine-tuned using the raw training texts over a limited number of epochs. This process allowed the model to learn task-specific patterns while leveraging its existing linguistic knowledge

All three models were then evaluated on the unseen testing set. Performance was measured using a confusion matrix to calculate four standard metrics: Accuracy, Precision, Recall, and F-1 Score [20]. The calculation formula of the those four standard metics is presented as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - Score = \frac{2\ X\ Precision\ X\ Recall}{Precision + Recall}$$

C. Result and Discussion

The machine learning and transformer-based models were trained and tested, and their performance metrics were recorded. The results are summarized in Table 1.

Table 1. Performance Metrics of Classification Models

No	Model	*Accuracy*	*Precision*	*Recall*	*F-1 Score*
1	Multinomial Naive Bayes	84,2%	83,5%	85,1%	84,3%
2	Support Vector Machine	89,5%	88,9%	90,2%	89,5%
3	IndoBERT	94,1%	93,8%	94,5%	94,1%

Experiment shows varying results for the different classification models. We found that IndoBERT model gave better prediction accuracy than the other test methods. The IndoBERT model resulted in an average prediction accuracy of 94,1%. However, the SVM model also performs well with accuracy of 89,5%. The Multinomial Naïve Bayes model achieves the lowest accuracy at 84,2%. The superior performance of IndoBERT is likely due to its ability to understand the context of words in a sentence, which is a known limitation of "bag-of-words" based machine learning models [7]. The Multinomial Naïve Bayes model's performance is likely hindered by its feature independence assumption, which is often violated in natural language [21].

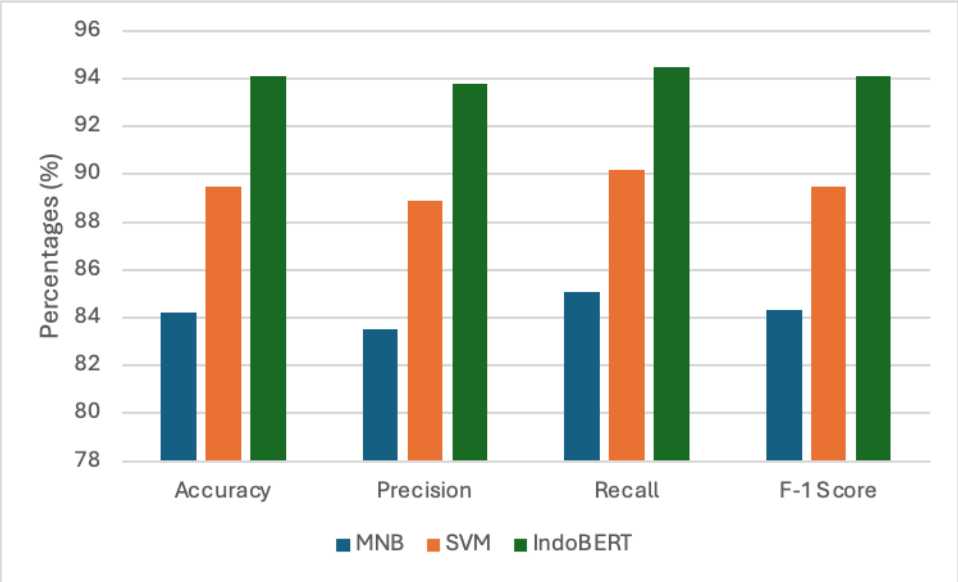

Figure 2. Visualization of Models Performance Metrics

Beyond overall accuracy of classification methods, the precision, recall, and F-1 score metrics provide a more nuanced understanding of each model's classification performance. Precisions measure the proportion of positively identified reviews that were correct, while recall measures the proportion of all actual positive reviews that were successfully found. The F-1 score provide a single, balanced measure by calculating the harmonic mean of precision and recall, which is particularly useful when evaluating the trade-off between the two [20].

As shown in Table 1, IndoBERT's high precision (93,8%) indicates it make very few false positive errors, meaning when it classifies reviews as positive, that

classification is highly reliable. Its high recall (94,5%) shows that it is also effective at identifying most of all positive reviews, just missing a very few reviews. The resulting F-1 score of 94,1% indicates that the model achieves a strong balance between precision and recall, reflecting its robustness in handling both false positives and negatives. Such balanced performance is particularly valuable in practical applications, where minimizing both types of errors is critical. Based on the bar chart in Figure 2, it clearly shows that the transformer-based model, IndoBERT, outperformed other tested machine learning models across all four-performance metrics: accuracy, precision, recall, and F-1 score.

To gain deeper insight into the differences in model performance, we conducted an error analysis on a subset of misclassified reviews from the test set. This analysis focuses on intances where the SVM model failed to correctly classify the sentiment, while the IndoBERT model succeeded. The goal was to highlight the specific linguistic challenges that may explain the performance gap between the models. Table 2 presents representative examples of these misclassifications.

Table 2. Example of Misclassfication Reviews

No	Review Text (Indonesian)	*Actual Sentiment*	*SVM Prediction*	*IndoBERT Prediction*	*Linguistic Challenge*
1	"Packingnya sih aman, tapi pengiriman tidak secepat yang diharapkan" English: "*The packaging was secure, but the delivery was not as fast as excepted*"	Negative	Positive	Negative	**Contrasting Clause:** The model focused on "aman" (secure) and missed the negative sentiment introduced by "tapi" (but)
2	"Barangnya tidak jelek, sesuai harga lah" English: "*The item is not bad, it is reasonable for the price*"	Positive	Negative	Positive	**Negation:** The presence of "jelek" (bad) misled the model, which failed to process the negation "tidak" (not)
3	"Kirain bakal kecewa, ternyata produknya bagus banget!" English: "*I thought I would be disappointed, turns out the product turned out to be really great*"	Positive	Negative	Positive	**Reversal of Expectation:** The model incorrectly weighted "kecewa" (dissappointed) without understanding the context reversal from "ternyata" (turns out)

Table 2 highlights a key limitation of the bag-of-words approach used by the SVM model. The model struggles with linguistic phenomena where contextual meaning is critical. In the case of negation and contrasting clauses, the SVM model incorrectly weight negative keywords like "jelek" or "kecewa" without considering the surrounding words ("tidak", "tapi", "ternyata") that alter or reverse the overall sentiment of the sentence. In contrast, the IndoBERT model with its self-attention

mechanism, is able to interpret these complex relationships between words, leading to a more accurate classification. This ability to understand context explains the significant performance advantage of transformer-based model like IndoBERT over machine learning models.

Table 3. Computational Performance Comparison

No	Model	*Training Time on ENV1*	*Training Time on ENV2*
1	Multinomial Naive Bayes	0,5s	0,5s
2	Support Vector Machine	2,3s	2,1s
3	IndoBERT	10812s	1805s

The data in Table 3 clearly illustrates the computational limitations of transformer-based models on standard hardware. We tested both machine learning and transformer-based models in two different hardware environments. The first environment is using Intel Core i7-12400H, and second environment is Nvidia Tesla T4 via Google Colab. On high performance CPU (ENV1), the Multinomial Naïve Bayes and SVM models train in seconds. In contrast, the IndoBERT model is computationally impractical, taking approximately 3 hours to complete the same task. This demonstrates that while transformer models are architecturally capable of running on CPU's, they are not designed for it.

However, when moved into its optimal hardware environment (ENV2), the IndoBERT model's training time is significantly reduced to 30 minutes. Importantly, the performance of the Multinomial Naïve Bayes and SVM does not improve in GPU environment, as these algorithms are not designed for parallel processing and still rely on the CPU. This highlights the necessity of specialized hardware for the practical deployment of large deep learning models. For business requiring the highest possible accuracy, the investment in GPU resources for a model like IndoBERT is justified. Instead, for less critical applications or in resource-constrained environments, SVM provides a highly effective balance of predictive performance and computational efficiency.

D. Conclusion

The result of this study shows that the fine-tuned IndoBERT which is transformer-based model is suitable for handling the classification of sentiment analyisis on Indonesian e-commerce reviews. This study also presents a comparison with other classification algorithm in machine learning models, such as Multinomial Naïve Bayes and SVM in terms of predictive accuracy, precision, recall, and F-1 score to find the best method to solve text classification problem. All models are capable of classifying sentiment with a reasonable degree of performance metrics. However, the IndoBERT model showed significantly better performance, achieving accuracy of 94,1%, precision of 93,8%, recall of 94,5%, and F-1 score of 94,1%.

In terms of computational cost, the machine learning model has better performance than the transformer-based model. This presents a critical trade-off between performance and efficiency. For applications where high accuracy is critical, investing in GPU hardware is justified. However, in settings with limited

resources or less demanding requirements, machine learning models like SVM provides a practical balance between performance and computational cost.

E. Acknowledgment

The authors would like to express their sincere gratitude to the State Polytechnic of Malang for providing facilities and support necessary for the completion of this research.

F. References

[1] S. Verma, R. Sharma, S. Deb, and D. Maitra, "Artificial intelligence in marketing: Systematic review and future research direction," International Journal of Information Management Data Insights, vol. 1, no. 1, 2021, doi: 10.1016/j.jjimei.2020.100002.

[2] A. Rosário and R. Raimundo, "Consumer marketing strategy and e-commerce in the last decade: A literature review," Journal of Theoretical and Applied Electronic Commerce Research, vol. 16, no. 7, pp. 3003–3024, 2021, doi: 10.3390/jtaer16070164.

[3] A. M. Rahman, W. R. Amelia, F. A. Nasution, and Z. Zulham, "the Influence of Online Customer Review and Online Customer Rating on Purchase Decisions At Tokopedia (Case Study of Tokopedia Users in Medan District, Johor)," Dharmawangsa: International Journal of the Social Sciences, Education and Humanitis, vol. 3, no. 1, pp. 23–33, 2022, doi: 10.46576/ijsseh.v3i1.2975.

[4] H. Atsqalani, N. Hayatin, and C. S. K. Aditya, "Sentiment Analysis from Indonesian Twitter Data Using Support Vector Machine And Query Expansion Ranking," Jurnal Online Informatika, vol. 7, no. 1, pp. 116–122, 2022, doi: 10.15575/join.v7i1.669.

[5] R. F. Ananda, A. Syahri, and F. N. Hasan, "Sentiment Analysis of Customer Satisfaction in Gojek and Grab Application Reviews Using the Naive Bayes Algorithm," Jurnal Teknik Informatika (Jutif), vol. 5, no. 1, pp. 233–241, 2024, doi: 10.52436/1.jutif.2024.5.1.1680.

[6] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1–40, 2021, [Online]. Available: http://arxiv.org/abs/2004.03705

[7] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," Trans Assoc Comput Linguist, vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.

[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[9] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif Intell Rev, vol. 53, no. 6, pp. 4335–4385, 2020, doi: 10.1007/s10462-019-09794-5.

[10] T. Dhamija, Anjum, and R. Katarya, "Comparative Analysis of Machine Learning and Deep Learning Algorithms for Detection of Online Hate Speech,"

Lecture Notes in Mechanical Engineering, pp. 509–520, 2021, doi: 10.1007/978-981-16-0942-8_48.
[11] J. Philip, B. Veerasekharreddy, M. Harshini, I. V. S. L. Haritha, S. Patil, and S. K. Shareef, "A Comparative Study of Text Classification using Selective Machine Learning Algorithms," Proceedings of the 7th International Conference on Intelligent Computing and Control Systems, ICICCS 2023, pp. 482–484, 2023, doi: 10.1109/ICICCS56967.2023.10142474.
[12] J. Y. Tan, A. S. K. Chow, and C. W. Tan, "A Comparative Study of Machine Learning Algorithms for Sentiment Analysis of Game Reviews," The Journal of The Institution of Engineers, Malaysia, vol. 82, no. 3, 2022, doi: 10.54552/v82i3.101.
[13] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Naacl-Hlt 2019, 2018, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423.pdf
[14] S. Aras, M. Yusuf, R. Y. Ruimassa, E. A. B. Wambrauw, and E. B. Pala'langan, "Sentiment Analysis on Shopee Product Reviews Using IndoBERT," Journal of Information Systems and Informatics, vol. 6, no. 3, pp. 1616–1627, 2024, doi: 10.51519/journalisi.v6i3.814.
[15] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2024, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
[16] T. Ghazi, Indonesian Marketplace Product Reviews. 2022. [Online]. Available: https://www.kaggle.com/datasets/taqiyyaghazi/indonesian-marketplace-product-reviews
[17] D. Jurafsky and J. H. Martin, Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition, no. 3rd. 2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/
[18] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7649 LNCS, no. PART 1, pp. 508–524, 2012, doi: 10.1007/978-3-642-35176-1_32.
[19] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian," ACM Transactions on Asian Language Information Processing, vol. 6, no. 4, pp. 1–33, 2007, doi: 10.1145/1316457.1316459.
[20] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, [Online]. Available: http://arxiv.org/abs/2010.16061
[21] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," Soft comput, vol. 25, no. 3, pp. 2277–2293, 2021, doi: 10.1007/s00500-020-05297-6.