JITEEHA: Journal of Information Technology Application in Education, Economy, Health and Agriculture
Vol. 03, No. 2, June 2026, pp. 29~35, e-ISSN: 3090-3939
https://journal.iteeacademy.org/index.php/jiteeha/

DistilBERT-Based E-Commerce Sentiment Analysis

Zahri Aksa Dautd[1], Aviv Yuniar Rahman[2], Fitri Marisa[3]
[1]Department of Informatics Engineering, University of Widyagama Malang, Jl. Borobudur No. 35 Malang, Indonesia
[2]Department of Informatics Engineering, University of Widyagama Malang, Jl. Borobudur No. 35 Malang, Indonesia
[3]Department of Informatics Engineering, University of Widyagama Malang, Jl. Borobudur No. 35 Malang, Indonesia
Corresponding author: zahriaksa@gmail.com

Article Info

Article history:

Received February 26, 2026
Revised March 24, 2026
Accepted May 25, 2026

Keywords:

Sentiment Analysis
E-Commerce
Shopee
DistilBERT
Transformer

ABSTRACT

The rapid advancement of digital technology has driven significant growth in Indonesia's e-commerce sector, with Shopee emerging as one of the largest platforms generating millions of product reviews daily. These reviews contain valuable consumer opinions that can be analyzed to assess customer satisfaction, yet their massive volume makes manual analysis inefficient and subjective. This study aims to develop an automated sentiment analysis model using DistilBERT to classify Shopee product reviews into positive and negative sentiments. The dataset comprises approximately 1 million English-language reviews covering various product categories, including electronics, fashion, beauty, and household items. The research methodology involves text preprocessing, tokenization using DistilBertTokenizerFast, and fine-tuning of the DistilBERT model under multiple data-split ratios (90:10, 80:20, 70:30, 60:40). Experimental results demonstrate that DistilBERT achieved the highest accuracy of 94.8%, outperforming baseline models such as Naïve Bayes (88.4%) and SVM (89.6%). These findings confirm that DistilBERT effectively maintains a balance between accuracy, precision, and recall while offering high computational efficiency. This research contributes both methodologically and practically by establishing DistilBERT as a scientifically robust and resource-efficient solution for large-scale sentiment analysis in Indonesia's e-commerce environment.

Corresponding Author:

Zahri Dautd
Department of Informatics Engineering
University of Widyagama Malang
Jl. Borobudur No. 35 Malang, Indonesia
Email: zahriaksa@gmail.com

1. INTRODUCTION

The The development of digital technology has transformed the way people interact in economic activities, particularly through e-commerce platforms [1]. In Indonesia, Shopee has become one of the largest platforms, with millions of active users and a transaction volume that continues to increase each year. This rapid growth has given rise to a new phenomenon in the form of an explosion in the number of product reviews from consumers, covering various categories such as electronics, clothing, beauty products, and household necessities [2]. These reviews play an important role as indicators of customer satisfaction and seller service quality, as well as a source of information for potential buyers in making purchasing decisions [3]. However, the high volume of review data on platforms like Shopee makes manual analysis inefficient and prone to bias [4]. Therefore, an automated approach based on machine learning is needed to interpret consumer opinions quickly and accurately [5].

The main challenge in analyzing e-commerce reviews lies in extracting emotional information and determining consumer sentiment efficiently without losing contextual meaning [6]. Conventional methods such as Naïve Bayes and Support Vector Machine (SVM) based on TF-IDF features can indeed classify text, but they

are limited in their ability to understand semantic relationships between words [7]. As a result, classification outcomes are often less accurate when dealing with ambiguous or complex sentences [8]. This situation creates the need for sentiment analysis methods that are not only accurate but also capable of deeply understanding natural language context and operating efficiently at large data scales, such as Shopee product reviews [9].

Various previous studies have applied machine learning and deep learning approaches to e-commerce sentiment analysis [10]. Classical methods such as Naïve Bayes and Random Forest have achieved accuracies ranging from 72–90%, while transformer-based models such as BERT and IndoBERT are capable of reaching accuracies of up to 93–95% [11]. However, research that focuses on optimizing lightweight transformer models such as DistilBERT remains very limited. Most existing studies emphasize full BERT models or local variants like IndoBERT, which require substantial computational resources [12]. Therefore, a research gap exists to explore the extent to which lightweight models like DistilBERT can deliver performance comparable to larger models while offering higher computational efficiency [13].

This study proposes the use of DistilBERT as the primary model for sentiment analysis of Shopee product reviews [14]. Conceptually, DistilBERT is not merely a "lightweight" version of BERT, but the result of a knowledge distillation process that transfers behavioral knowledge from a full BERT model to a new model with an optimized structure [14]. Unlike other transformer models that simply reduce parameters or modify tokenization (such as RoBERTa or IndoBERT), DistilBERT introduces three key scientific innovations that make it well suited for large-scale sentiment analysis tasks :

a. First, DistilBERT retains BERT's bidirectional contextual embeddings but removes token-type embeddings that are only relevant for sentence-pair tasks (such as next sentence prediction). This design allows DistilBERT to focus more on intra-sentence context, which is highly suitable for single-sentence sentiment classification [15].

b. Second, the distillation process in DistilBERT does not merely prune layers, but optimizes the soft target probability distributions (output logits) of BERT to train the student model [16]. As a result, DistilBERT learns the generalization behavior and semantic smoothness of the teacher model, rather than simply mimicking word representations. This enables it to be more robust to variations in opinion expressions, sarcasm, and subtle sentence nuances phenomena that frequently appear in e-commerce user reviews [17].

c. Third, DistilBERT employs a triple-loss combination: (1) language modeling loss, (2) cosine embedding loss between the teacher and student representations, and (3) Kullback–Leibler divergence loss between their prediction distributions. This combination enables deeper semantic transfer compared to other models such as RoBERTa, which focuses solely on masked language modeling, or IndoBERT, which typically performs fine-tuning without a dedicated distillation stage [18].

With these characteristics, DistilBERT is not only more efficient but also more stable in capturing emotional context from short to medium-length texts the types of text that dominate Shopee reviews [19]. Therefore, the selection of DistilBERT in this study is not merely a computational consideration, but is grounded in its scientific advantages in semantic learning mechanisms, which make it highly suitable for context-rich and expressive natural language based sentiment analysis [20].

2. METHOD

This study employs a quantitative experimental approach aimed at evaluating the performance of the DistilBERT model in conducting sentiment analysis on Shopee product reviews. The experimental approach is chosen because it allows for direct testing of the model's effectiveness through training (fine-tuning) and testing processes using real-world datasets. In this study, the independent variable is the DistilBERT model used as the primary analytical method, while the dependent variable is the sentiment classification outcome in the form of positive or negative categories, measured using four evaluation metrics: accuracy, precision, recall, and F1-score.

The dataset used in this study is the Shopee Text Review Dataset obtained from the Kaggle platform. This dataset contains approximately one million English-language product reviews, each labeled with positive or negative sentiment. After the data cleaning process, the dataset was converted into .xlsx format to facilitate processing using Python. To maintain class balance, 47,430 positive reviews and 47,430 negative reviews were selected. The dataset was then split into training and testing sets using four different ratios 90:10, 80:20, 70:30, and 60:40 to observe the consistency of the model's performance across different data proportions.

Text preprocessing was carried out to ensure that the data were optimally prepared for processing by the transformer model. The process began with case folding (converting all text to lowercase), removal of non-alphabetic characters, and mapping sentiment labels to numerical values (1 for positive and 0 for negative). The data were split into training and testing sets using the train_test_split() function from the scikit-learn library to maintain class balance. Subsequently, tokenization was performed using DistilBertTokenizerFast with parameters max_length = 128, padding = True, and truncation = True to ensure uniform text length. The dataset was then wrapped into a PyTorch-based ShopeeDataset class so that it could be utilized in the DistilBERT model training process.

The primary model used in this study is DistilBERT-base-uncased, a product of knowledge distillation from BERT. This model was fine-tuned for binary classification using a batch size of 16, a learning rate of 5e-5, and the AdamW optimizer (Adam with Weight Decay), with the number of training epochs varied from 1 to 10. DistilBERT was selected because it retains approximately 97% of BERT's performance while using about 40% fewer parameters, and it incorporates a triple-loss mechanism (a combination of language modeling loss, cosine embedding loss, and Kullback–Leibler divergence loss), which enhances both computational efficiency and semantic understanding of the text.

As benchmarks, two baseline models were employed: Naïve Bayes with TF-IDF and SVM with TF-IDF. Both were used to measure the extent of performance improvement achieved by DistilBERT. Naïve Bayes relies on word probability distributions across classes, while SVM applies hyperplane optimization based on TF-IDF–weighted word features.

Evaluation was conducted using four main metrics: accuracy, precision, recall, and F1-score. These four metrics were calculated based on the confusion matrix results, which represent the number of correct and incorrect predictions for each class. The general structure of the confusion matrix for binary classification is shown in Table 1.

Table 1. Confusion Matrix

Actual Class	Positive Prediction	Negative Prediction
Positive	True Positive (TP)	False Negative (FP)
Negative	False Positive (FP)	True Negative (TP)

From this matrix, the evaluation metrics are calculated using the following equation:

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

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

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

$$F1 = 2X\frac{Precission \times Recall}{Precision + Recall}$$

Model evaluation was conducted using four main metrics: accuracy, precision, recall, and F1-score. These metrics were calculated based on the confusion matrix, which represents the number of correct and incorrect predictions for each class. The general structure of the confusion matrix for binary classification is presented in Table 1.

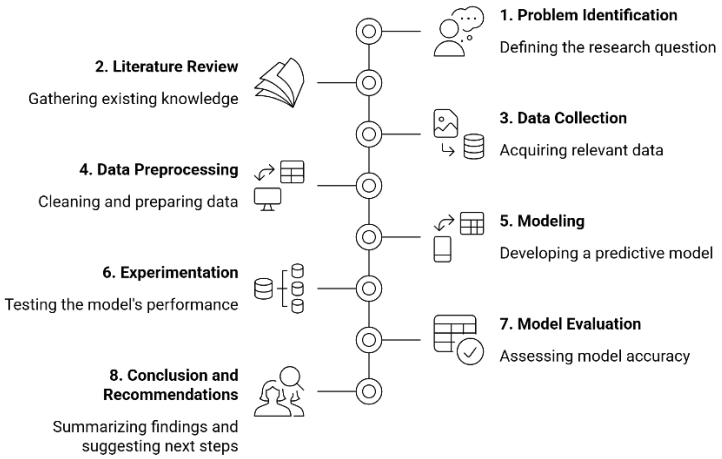
Confusion matrix analysis provides a comprehensive overview of prediction error patterns. For instance, a high False Positive (FP) value indicates that the model tends to be overly sensitive to negative reviews (over-classification), while a high False Negative (FN) value suggests that the model fails to correctly identify positive reviews that should have been classified as such. Therefore, this analysis is important for understanding not only how accurate the model is, but also how it behaves when dealing with ambiguous data or reviews containing mixed sentiments.

Overall, the research workflow begins with the data collection stage, which involves downloading and preparing the Shopee review dataset from the Kaggle platform. The next stage is data preprocessing, which includes text cleaning, tokenization, and the construction of structured data.

This is followed by the model training stage, where the DistilBERT model is fine-tuned and two baseline models are trained for comparative purposes. The subsequent stage is model testing to obtain sentiment prediction results on the test data, which are then evaluated using performance metrics. After all results are obtained, a comparative analysis is conducted between the performance of DistilBERT and the baseline models to assess the extent of performance improvement achieved by the lightweight transformer model.

The study concludes with the formulation of scientific conclusions and a discussion of the practical implications of using DistilBERT in sentiment analysis systems for e-commerce platforms. Visually, the entire research process is illustrated in the form of a flow diagram, as shown in Figure 1.

Figurre 1. Flow Diagram



3. RESULTS AND DISCUSSION

This study produced a sentiment analysis model based on DistilBERT, which was then compared with two baseline models: Naïve Bayes with TF-IDF and SVM with TF-IDF. The experiments were conducted using four data split ratios 90:10, 80:20, 70:30, and 60:40 to evaluate the stability and consistency of model performance across different proportions of training and testing data.

The experimental results show that DistilBERT consistently achieved the best performance across all data ratios. The highest accuracy was obtained at the 90:10 split, reaching 94.8%, with precision and recall values of 94.5% and 94.3%, respectively. The model's performance remained stable at the 80:20 and 70:30 ratios, achieving accuracies of 94.5% and 94.2%. At the 60:40 ratio, accuracy slightly decreased to 93.7% due to the reduced proportion of training data.

In contrast, the two baseline models demonstrated lower performance. The Naïve Bayes model achieved a maximum accuracy of only 88.4%, while the SVM model reached 89.6% accuracy at its best ratio. Overall, DistilBERT outperformed the conventional methods by an average margin of 5–6%, while also exhibiting superior performance stability across all data ratio scenarios.

Table 2 below displays the complete test results of each model on the data ratio variations:

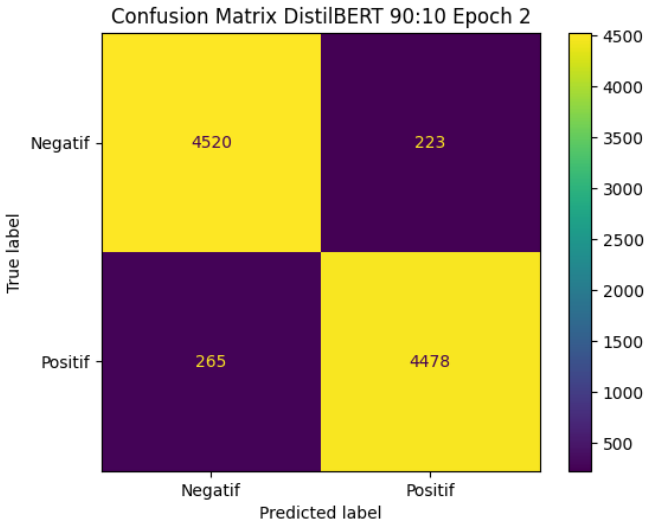
Table 2. Complete Test Result

Model	Rasio Data (Train: Test)	Accuracy	Precission	Recall	F1-Score
DistilBERT	90:10	0.94	0.94	0.94	0.94
DistilBERT	80:20	0.94	0.94	0.94	0.94
DistilBERT	70:30	0.94	0.93	0.93	0.93
DistilBERT	60:40	0.93	0.93	0.93	0.93
Naive Bayes + TF-IDF	80:20	0.88	0.88	0.87	0.87
SVM + TF-IDF	80:20	0.89	0.89	0.89	0.89

Tables Based on the table above, DistilBERT not only outperforms the other models in terms of accuracy, but also demonstrates a high balance between precision and recall, resulting in the highest F1-score among all models. This indicates that the model is able to consistently identify both positive and negative sentiments with a low error rate. To examine the distribution of predictions, a confusion matrix is applied to the best-performing model, namely DistilBERT with a 90:10 data split ratio. The results of the confusion matrix are presented in Figure 2.

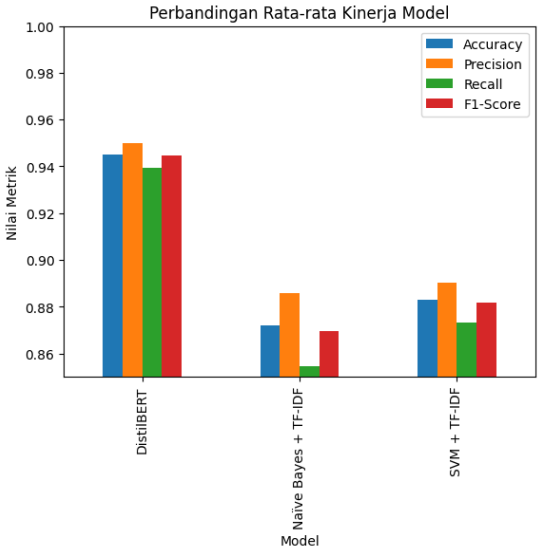
From the confusion matrix, it can be seen that the True Positive (TP) and True Negative (TN) values dominate compared to the misclassification values (False Positive and False Negative). This indicates that DistilBERT is able to recognize sentiment patterns with a prediction error rate below 3%. This low error proportion reinforces the evaluation metrics results, which show a balanced performance between precision and recall.

Figure 2. 90:10 data split ratio



To further clarify the performance comparison among the models, a bar chart was created, as shown in Figure 3, illustrating the accuracy comparison across the three models.

Figure 3. Accuracy Comparison



DistilBERT's ability to maintain a balance between True Positives (TP) and True Negatives (TN) serves as a strong indicator that the model can effectively understand sentence context [21]. For example, in ambiguous sentences such as "the product is good but delivery is slow," the model can accurately identify the main emotional polarity, unlike Naïve Bayes and SVM, which tend to misclassify due to their limitations in capturing semantic relationships between words. The experimental results above demonstrate that the transformer-based DistilBERT model has far superior semantic representation capabilities compared to classical word-weighting methods like TF-IDF. Through its bidirectional self-attention mechanism and knowledge distillation process, DistilBERT can learn the distribution of meaning across words in a sentence, making it more sensitive to emotional context and syntactic structure [22].

This capability allows the model to recognize subtler sentiment nuances, such as differentiating between "the product is good but delivery is slow" and "the product is bad but delivery is fast," which conventional models often classify as neutral. This explains why DistilBERT exhibits lower false positive and false negative rates. Compared to previous studies, for example, IndoBERT by Aras et al. (2024) with 93% accuracy, and RNN-QER by Novitasari et al. (2024) with 95% accuracy, this study shows that DistilBERT can achieve performance comparable to larger models while being more computationally efficient. This advantage is attributed to the parameter distillation process, where the "student" model learns from the probability distributions of the "teacher" model, thereby maintaining deep semantic understanding without high complexity.

Furthermore, testing shows that optimal fine-tuning is achieved between the 2nd and 3rd epochs, after which the model's performance tends to stabilize. This indicates a fast convergence rate, making DistilBERT more efficient in training time compared to full transformer models like BERT. Thus, these results not only confirm DistilBERT's superior performance but also demonstrate an optimal trade-off between training speed, accuracy, and computational resource efficiency, making it an ideal model for large-scale data processing scenarios such as e-commerce reviews.

This study provides significant scientific contributions compared to previous research, both methodologically, technically, empirically, and practically. Methodologically, it is among the first studies to apply DistilBERT for large-scale sentiment analysis of Shopee reviews. Unlike previous studies that used BERT or IndoBERT on limited datasets, this research systematically tests the stability of a lightweight model across different data split ratios to observe consistency in performance. Technically, DistilBERT is optimized through efficient fine-tuning configurations with a batch size of 16, learning rate of 5e-5, and 1–10 epochs, achieving high performance even in the early epochs. This demonstrates that the model can reach high accuracy without requiring extensive computational resources, providing a cost-effective yet precise solution for large-scale text analysis.

Empirically, the testing results show that DistilBERT delivers an average accuracy improvement of 5–6% compared to classical methods such as Naïve Bayes and SVM, while also achieving better balance between precision and recall across all data split ratios. These findings confirm that DistilBERT is adaptive to the complexities of natural language and varying emotional contexts in consumer reviews. Practically, this study opens opportunities for applying DistilBERT in automated sentiment analysis systems on e-commerce platforms to understand public perception, improve service quality, and support data-driven decision-making.

In summary, this research demonstrates that DistilBERT is not merely a lightweight version of BERT, but a scientifically grounded solution that balances accuracy, efficiency, and deep semantic understanding making it an ideal model for large-scale sentiment analysis systems in Indonesia's e-commerce industry.

4. CONCLUSION

The This study successfully demonstrates that DistilBERT is a lightweight transformer model capable of delivering high performance in sentiment analysis of e-commerce product reviews, particularly on the Shopee platform. The experimental results show that DistilBERT consistently achieves the highest accuracy, reaching 94.8% with a 90:10 data split, along with balanced precision and recall values above 94%. These findings confirm that the model can accurately identify positive and negative sentiment patterns with a low prediction error rate, even across different data split ratios. Compared to the two baseline models Naïve Bayes and SVM with TF-IDF DistilBERT outperforms by an average of 5–6% across all main evaluation metrics.

Methodologically, this study is among the first to apply DistilBERT to large-scale Shopee review datasets while also testing the model's stability across four different data split ratios. Technically, the study demonstrates that an efficient fine-tuning configuration with a batch size of 16, a learning rate of 5e-5, and epochs ranging from 1 to 10 can achieve high performance even in early epochs, significantly reducing training time and computational resource consumption. Empirically, DistilBERT proves to have the best balance among accuracy, precision, recall, and F1-score, with a stronger semantic representation capability compared to conventional word-weighting methods.

Furthermore, the practical contribution of this study is the demonstration that DistilBERT can serve as the foundation for automated sentiment analysis systems in Indonesia's e-commerce sector, supporting tasks such as monitoring customer opinions, product recommendation systems, and real-time customer satisfaction detection. The model can be implemented in production systems without requiring large-scale computational infrastructure, making it an ideal solution for companies with limited resources.

In conclusion, this study confirms that DistilBERT is not merely a lightweight version of BERT, but a scientifically grounded approach that balances accuracy, efficiency, and deep semantic understanding, positioning it as a potential standard model for large-scale sentiment analysis in the future.

REFERENCES

[1] A. P. Lestari, S. A. Fatiha, and S. O. Putri, "International Journal of Computer in Law & Political Science E-Commerce in Indonesia ' s Economic Transformation and Its Influence on Global Trade," vol. 4, pp. 10–23, 2024.

[2] S. E-commerce and E. Shopee, "International Journal Administration, Business & Organization," vol. 5, no. 3, pp. 118–128, 2024.

[3] L. Hartimar, Y. Manza, and K. P. Siregar, "Text Classification Using TF-IDF and Naïve Bayes : Case Study of MyXL App User Review Data," vol. 2, no. 2, pp. 100–108, 2025.

[4] D. Alvionita and J. Parhusip, "Analisis Feedback Pengguna Aplikasi Shopee Menggunakan Distribusi Proporsi dan Teknologi Informasi," vol. 5, pp. 13–17, 2025.

[5] K. Neighbor, "Komparasi Efektifitas Analisis Sentimen pada Ulasan Produk E-Commerce Menggunakan Naive Bayes," vol. 7, no. 2, pp. 226–236, 2025.

[6] Ramadila et al, "SENTIMENT DETECTION OF SHOPEE E-COMMERCE APPLICATION REVIEWS USING

NATURAL LANGUAGE PROCESSING AND SUPPORT VECTOR MACHINE," vol. 1, pp. 28–38, 2025.

[7] A. Salsabila, B. Priyatna, and A. Hananto, "Komparasi Kinerja Model Naive Bayes , SVM , dan BERT dalam Klasifikasi Sentimen Ulasan Pada Aplikasi YUMMY," vol. 4, no. 2, pp. 42–47, 2025.

[8] A. A. Solihin *et al.*, "Evaluasi Pengaruh Varian Daftar Stopword terhadap Kinerja Klasifikasi Teks Al- Qur ' an dengan Support Vector Machine dan Backpropagation Neural Network Program Studi Teknologi Informasi , Fakultas Ilmu Komputer , Universitas Amikom Purwokerto , Indonesia Evaluation of the Impact of Stopword List Variants on Quranic Text Classification Performance Using Support Vector Machine and Backpropagation Neural Network," vol. 5, no. 7, pp. 1867–1880, 2025.

[9] M. Souppaya *et al.*, "Analisis Sentimen Pada Ulasan 'Lazada ' Berbahasa Indonesia Menggunakan K-Nearest Neighbor (K-NN) Dengan Perbaikan Kata Menggunakan Jaro Winkler Distance," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, 2022.

[10] U. Gunadarma, "Sentiment analysis methods for customer review of indonesia e-commerce," vol. 20, no. 1, pp. 47–60, 2024, doi: 10.24507/ijicic.20.01.47.

[11] V. No, J. Hal, A. Kamal, and R. Astri, "Eksplorasi Sentimen Pengguna pada Aplikasi E-Commerce dengan Deep Learning," vol. 7, no. 3, pp. 435–441, 2025.

[12] E. Abdurachman and L. A. Wulandhari, "Development of a Model to Detect the Validity of Indonesian Reviews on E-Commerce Products Using Bert and Smart Approaches," pp. 2287–2295, 2024.

[13] R. F. Herdiyanto and M. Thoriq, "Sentiment Analysis of Marketplace Review with Islamic Perspective using Fine-Tuning DistilBERT," vol. 2, no. 2, 2025.

[14] W. Christian, D. Adamlu, A. Yu, and D. Suhartono, "Leveraging IndoBERT and DistilBERT for Indonesian Emotion Classification in E-Commerce Reviews arXiv : 2509 . 14611v1 [cs . CL] 18 Sep 2025," vol. 00, 2025.

[15] F. Fajri, B. Tutuko, and S. Sukemi, "Membandingkan Nilai Akurasi BERT dan DistilBERT pada Dataset Twitter Tahapan Penelitian," vol. 8, no. 2, 2022.

[16] A. Novitasari, Y. Sibaroni, and D. Puspandari, "Multi-aspect Sentiment Analysis of Shopee Application Reviews using RNN Method and Query Expansion Ranking," *Build. Informatics, Technol. Sci.*, vol. 6, no. 2, pp. 825–834, 2024, doi: 10.47065/bits.v6i2.5605.

[17] Najwa Fathiro Cahyono, Khurrotul 'Uyun, and Siti Mukaromah, "ETIKA PENGGUNAAN KECERDASAN BUATAN PADA TEKNOLOGI INFORMASI," *Pros. Semin. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 1, 2023, doi: 10.33005/sitasi.v3i1.334.

[18] M. Hussain *et al.*, "Optimised knowledge distillation for efficient social media emotion recognition using DistilBERT and," pp. 1–16, 2025.

[19] H. B. Firmansyah, A. Afriansyah, and V. Lorini, "Comparing BERTBase , DistilBERT and RoBERTa in Sentiment Analysis for Disaster Response," vol. 6, no. 5, pp. 3419–3429, 2025.

[20] K. Piasta, "Comparative Analysis of Natural Language Processing Techniques in the Classification of Press Articles," 2025.

[21] M. I. Fahmi, A. A. Nababan, R. Optimized, and B. Pretraining, "Mathematical Modelling of Engineering Problems A Comparative Study of BERT and RoBERTa for Sentiment Analysis on Twitter Data Related to Mental Health," vol. 12, no. 9, pp. 3289–3295, 2025.

[22] A. Awalina, F. A. Bachtiar, F. Utaminingrum, U. Brawijaya, and P. Korespondensi, "PERBANDINGAN PRETRAINED MODEL TRANSFORMER PADA DETEKSI COMPARISON OF PRETRAINED TRANSFORMER MODELS ON SPAM REVIEW," vol. 9, no. 3, 2022, doi: 10.25126/jtiik.202295696.