Journal of Informatics and Vocational Education (JOIVE)
Vol.9, No.3, November 2026, pp. 67~77
ISSN: 2746-7813

Sentiment Analysis and Topic Modeling of Ruangguru Application User Reviews Using IndoBERT and BERTopic on a Kaggle Dataset

M. Fajar Ramadhan[1], Febrianti Panjaitan[2], Winarnie[3] , Hery Oktafiandi[4], Yohanes[5]
[1,2,3] Teknik Informatika, Fakultas Teknologi Kreatif, Satu University, Palembang
[4,5]Sistem Informasi, Fakultas Teknologi Kreatif, Satu University, Palembang

Article Info

Article history:

Received Jun 11, 2026
Revised Jun 18, 2026
Accepted Nov 1, 2026

Corresponding Author:

M. Fajar Ramadhan,
Satu University, Rukan Taman Harapan Indah B3 & B5, Jalan Letda Abdul Rozak Street, Palembang City, South Sumatera, 30114
Email:
m.ramadhan@univ.satu.ac.id

ABSTRACT

User reviews provide valuable information for evaluating digital learning platforms because they contain direct expressions of user satisfaction, complaints, and expectations. This study analyzed user reviews of the Ruangguru application using a Kaggle dataset by combining IndoBERT for sentiment classification and BERTopic for topic modeling. The study aimed to classify user sentiment, compare IndoBERT with a TF-IDF and Logistic Regression baseline, identify dominant discussion topics, and examine whether the model performance difference was statistically significant. The dataset originally contained 100,000 reviews, and 76,699 valid reviews were used after data cleaning and preprocessing. Sentiment labels were generated from rating values and grouped into negative, neutral, and positive classes. IndoBERT achieved an accuracy of 0.8920 and an F1 macro score of 0.6347, outperforming the baseline model with an accuracy of 0.8092 and an F1 macro score of 0.5666. McNemar's test confirmed that the performance difference was statistically significant (chi-square = 594.30, $p < 0.001$). BERTopic generated 41 topic groups, including one outlier group. Positive sentiment dominated most topics, particularly those related to learning support, material comprehension, and video-based learning. Negative sentiment was concentrated in topics related to payment, application updates, advertisements, and account issues. These findings show that integrating IndoBERT and BERTopic provides a comprehensive understanding of user perceptions toward educational applications.

Keywords: Sentiment Analysis, IndoBERT, BERTopic, Ruangguru, User Reviews

1. INTRODUCTION

The rapid expansion of educational technology has changed how students access learning materials, academic assistance, and digital tutoring services. In Indonesia, Ruangguru has become one of the most prominent online learning platforms by providing video-based lessons, practice questions, tutoring services, learning packages, and other educational features. As the number of users increases, user reviews become an important source for evaluating the quality of digital learning services because they contain textual opinions related to satisfaction, dissatisfaction, technical barriers, perceived usefulness, and user expectations.

User-generated reviews can be analyzed using natural language processing (NLP), particularly sentiment analysis. Sentiment analysis classifies textual data into polarity categories such as positive, neutral, and negative. Recent studies show that NLP-based sentiment analysis has developed rapidly through deep learning and Transformer-based models, although challenges remain in handling informal language, class imbalance, context dependency, and domain-specific expressions (Jim et al., 2024). In educational applications, sentiment analysis helps identify how users perceive learning content, application usability, pricing, accessibility, and technical performance. However, sentiment classification alone cannot fully explain the specific issues behind user opinions, so it should be complemented by topic modeling.

Research on Indonesian NLP has been strengthened by the development of pre-trained language models. Koto et al. (2020) introduced IndoBERT through the IndoLEM benchmark and showed that it achieved strong performance across several Indonesian language understanding tasks. Cahyawijaya et al. (2021) also provided benchmark resources for Indonesian natural language processing, supporting the evaluation of Indonesian language models. In addition, Koto et al. (2021) introduced IndoBERTweet, a domain-specific pre-trained language model for Indonesian Twitter data. These studies show the importance of contextual language representation for Indonesian informal and user-generated content.

Several previous studies have applied IndoBERT and related deep learning approaches for Indonesian sentiment analysis. Jayadianti et al. (2022) demonstrated that fine-tuning IndoBERT could produce strong performance in Indonesian review sentiment classification, while Geni et al. (2023) applied IndoBERT language models to Indonesian political tweets and found that Transformer-based models outperformed several traditional baselines. Other studies also explored CNN-GRU and feature expansion approaches for sentiment classification on Indonesian social media data (A. Z. R. Adam & Setiawan, 2023). In educational discourse, Wafda et al. (2025) applied aspect-based sentiment analysis using IndoBERT to Twitter discussions about the Merdeka Curriculum. However, the use of predefined aspects differs from the present study, which uses BERTopic to discover topics automatically from user reviews.

Although IndoBERT has shown strong performance in sentiment classification, sentiment should not be interpreted independently from topic context. Saputra et al. (2026) emphasized that sentiment classification may lose important meaning when ignoring the topic or context surrounding sentiment expressions. Similarly, studies comparing conventional machine learning, recurrent neural networks, and Transformer-based models show that model performance can vary depending on domain, text characteristics, and evaluation setting (Mayzaroh et al., 2026; Purwanto, 2026). Therefore, combining sentiment classification with topic-level interpretation is necessary to obtain a more comprehensive understanding of user opinions.

Topic modeling is useful for discovering hidden thematic structures in large-scale text data. BERTopic is a neural topic modeling approach that combines Transformer-based document embeddings, dimensionality reduction, clustering, and class-based TF-IDF to generate interpretable topic representations (Grootendorst, 2022). Compared with traditional topic modeling approaches such as LDA and NMF, BERTopic can capture semantic relationships more effectively because it relies on contextual embeddings rather than only word co-occurrence patterns. Egger & Yu (2022) showed that BERTopic could provide competitive and interpretable topic modeling results, while Ramamoorthy et al. (2024) demonstrated that topic modeling can be expanded by integrating sentiment classification.

The integration of sentiment analysis and topic modeling has increasingly been used to provide deeper insights into public opinion and user experience. Hanny and Resch (2024) proposed a joint topic-sentiment modeling approach, while Li and Hu (2025) applied BERTopic-based topic-sentiment collaborative analysis to compare scholarly and public perspectives on medical artificial intelligence. In the Indonesian context, Sihombing and Widiyaningtyas (2025) combined IndoBERTweet and BERTopic to analyze public health issues, and Anugrah & Agatha (2025) used IndoBERT and BERTopic to explore public opinion on BPS Instagram posts. Similar topic-sentiment integration has also been applied to user reviews and public service applications (Abel et al., 2026; N. P. Adam et al., 2026; Pinilih et al., 2026).

Based on the reviewed studies, there remains a research gap in the analysis of Indonesian educational application reviews, especially Ruangguru user reviews obtained from Kaggle. Previous studies have demonstrated the effectiveness of IndoBERT for sentiment classification and BERTopic for topic modeling, but limited research has combined both methods to analyze large-scale user reviews in the digital learning domain while statistically testing the difference between the baseline and Transformer-based classifier. Therefore, this study integrates IndoBERT, BERTopic, and McNemar's test to classify user sentiment, compare model performance, identify dominant topics, and map sentiment distribution across topics.

2. RESEARCH METHOD

This study employed a quantitative computational approach based on natural language processing. The research workflow consisted of dataset collection, data cleaning, sentiment labeling, text preprocessing,

baseline model development, IndoBERT fine-tuning, model evaluation, statistical significance testing, BERTopic modeling, and topic-sentiment mapping.

2.1. Dataset

The dataset used in this study was obtained from Kaggle and consisted of user reviews of the Ruangguru application. The original dataset contained 100,000 review records with four main attributes: userName, score, at, and content. The content attribute contained textual reviews written by users, while the score attribute represented the rating assigned by users to the application.

Before model development, the dataset was cleaned to remove incomplete, empty, and irrelevant review texts. After preprocessing, 76,699 valid review records were used for sentiment prediction and topic modeling. This dataset was considered suitable because it represented direct user experiences with Ruangguru, including satisfaction, criticism, complaints, and suggestions.

2.2. Sentiment Labeling

Sentiment labels were generated based on user rating values. Low ratings were categorized as negative sentiment, middle ratings were categorized as neutral sentiment, and high ratings were categorized as positive sentiment. This rating-based labeling approach was selected because user ratings in application reviews commonly reflect the overall polarity of user experience.

The sentiment classes consisted of three categories: negative, neutral, and positive. The negative class represented dissatisfaction, complaints, or unfavorable experiences. The neutral class represented moderate or ambiguous opinions. The positive class represented satisfaction, appreciation, or favorable experiences with the application.

2.3. Text Preprocessing

Text preprocessing was conducted to prepare the review data for model training and topic modeling. The preprocessing steps included removing missing values, cleaning irrelevant symbols, normalizing textual content, and preparing clean review text. This step was important because user-generated reviews often contain informal expressions, spelling variations, abbreviations, repeated characters, punctuation inconsistencies, and mixed writing styles.

For IndoBERT, the cleaned texts were processed using a tokenizer compatible with the selected IndoBERT model. Tokenization converted the review texts into token IDs and attention masks required by the Transformer architecture. For BERTopic, the cleaned review texts were used as document inputs to generate semantic embeddings and topic clusters.

2.4. Baseline Model: TF-IDF and Logistic Regression

A baseline model was developed using TF-IDF and Logistic Regression. TF-IDF transformed review texts into numerical feature vectors based on term frequency and inverse document frequency. Logistic Regression was then applied as a classifier to predict sentiment classes. The baseline model was included to compare a conventional machine learning approach with a contextual Transformer-based model.

2.5. IndoBERT Fine-Tuning

The main sentiment classification model used in this study was IndoBERT. IndoBERT was selected because it was specifically developed for Indonesian NLP tasks and can capture contextual relationships among words in Indonesian texts (Koto et al., 2020). The model was fine-tuned using the labeled Ruangguru review dataset, and the data were divided into training, validation, and test sets. During training, model performance was monitored using training loss, validation loss, validation accuracy, and validation F1 macro.

2.6. Model Evaluation

Model evaluation compared the baseline model and IndoBERT using accuracy, precision macro, recall macro, and F1 macro. Accuracy measured the proportion of correct predictions among all predictions. Precision macro measured the average precision across all sentiment classes, recall macro measured the

average ability of the model to identify each class, and F1 macro represented the harmonic mean of precision and recall across classes. Macro-averaged metrics were used because the sentiment classes were imbalanced.

2.7. Statistical Significance Testing

To examine whether the performance difference between TF-IDF + Logistic Regression and IndoBERT was statistically significant, this study applied McNemar's test. This test was selected because both classifiers were evaluated using the same test set, making their predictions paired at the instance level. McNemar's test evaluates the disagreement between two classifiers by comparing the number of instances correctly classified by the first model but incorrectly classified by the second model, and vice versa. A significance level of 0.05 was used in this study.

2.8. Topic Modeling Using BERTopic

After sentiment classification, BERTopic was applied to identify latent topics in user reviews. BERTopic generated document embeddings, reduced the dimensionality of the embedding space, clustered semantically similar documents, and represented topics using class-based TF-IDF (Grootendorst, 2022). The output included topic IDs, topic sizes, representative keywords, and representative documents. In this study, BERTopic generated 41 topic groups, including one outlier topic.

2.9. Topic Sentiment Mapping

The final stage involved mapping sentiment predictions to the topics generated by BERTopic. Each review had both a sentiment label and a topic label. By combining these two outputs, this study calculated the proportion of negative, neutral, and positive sentiment within each topic. This step provided a more detailed interpretation than sentiment analysis alone because it connected sentiment polarity with specific discussion themes.

3. RESULT AND ANALYSIS

3.1. RESULT

This section presents the experimental results obtained from sentiment classification using IndoBERT and topic modeling using BERTopic. The results include predicted sentiment distribution, IndoBERT training performance, model comparison, statistical significance testing, confusion matrix analysis, BERTopic topic distribution, and topic-sentiment mapping.

The initial dataset consisted of 100,000 Ruangguru user reviews collected from Kaggle. After data cleaning and preprocessing, 76,699 valid reviews were used for sentiment prediction and topic modeling. The predicted sentiment distribution obtained from the IndoBERT model is presented in Table 1 and Figure 1.

Table 1. Distribution of predicted sentiment labels

Sentiment	Count	Percentage
Positive	63,760	83.13%
Negative	10,281	13.40%
Neutral	2,658	3.47%
Total	76,699	100.00%

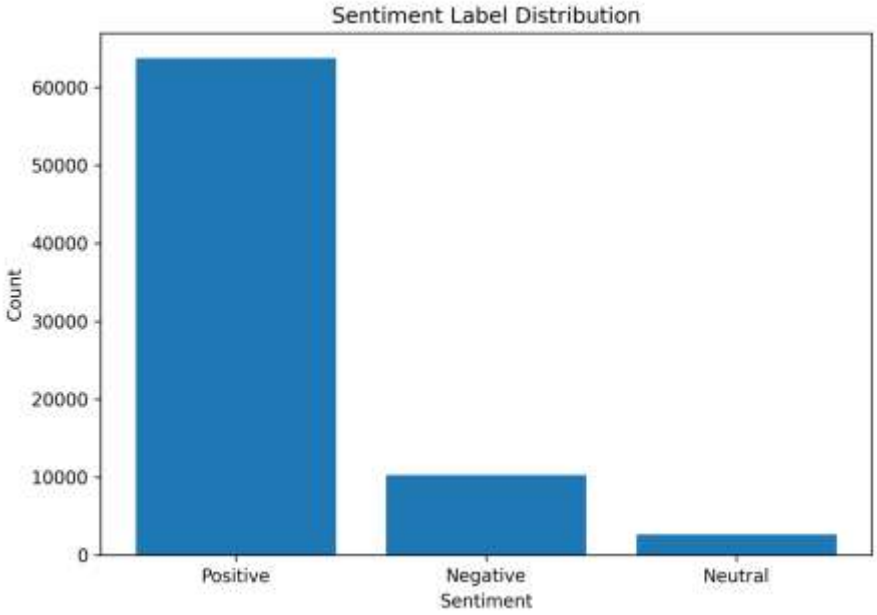

Figure 1. Distribution of predicted sentiment labels

Table 1 shows that positive sentiment dominated the predicted sentiment distribution, representing 83.13% of all valid reviews. Negative sentiment accounted for 13.40%, while neutral sentiment represented 3.47%. This distribution indicates that most Ruangguru users expressed favorable opinions about the application, although a smaller portion of reviews still contained dissatisfaction or ambiguous responses.

The IndoBERT training history is shown in Table 2.

Table 2. IndoBERT training history

Epoch	Training Loss	Validation Loss	Validation Accuracy	Validation F1 Macro
1	0.7754	0.7206	0.8787	0.6357
2	0.6618	0.8553	0.9024	0.6342
3	0.5412	1.0136	0.8966	0.6373

The training loss decreased from 0.7754 in the first epoch to 0.5412 in the third epoch, indicating that the model progressively learned sentiment patterns from the training data. The highest validation accuracy was achieved in the second epoch with a value of 0.9024, while the highest validation F1 macro was achieved in the third epoch with a value of 0.6373.

The performance comparison between the baseline model and IndoBERT is presented in Table 3 and Figure 2.

Table 3. Performance comparison between sentiment classification models

Model	Accuracy	Precision Macro	Recall Macro	F1 Macro
TF-IDF + Logistic Regression	0.8092	0.5379	0.6528	0.5666
IndoBERT	0.8920	0.6249	0.6460	0.6347

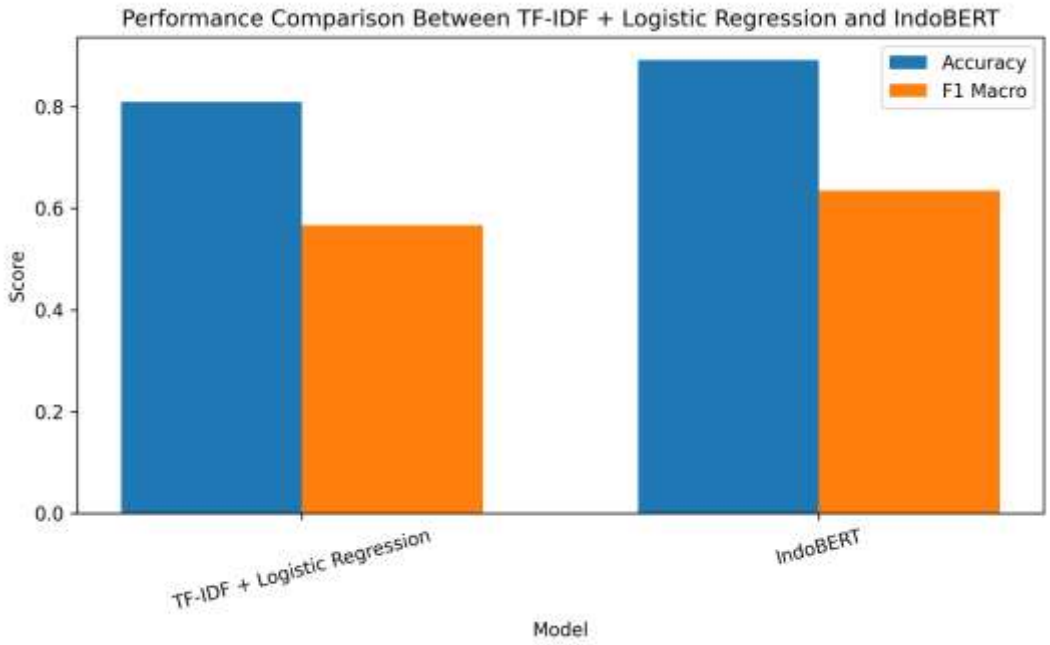

Figure 2Performance comparison between TF-IDF + Logistic Regression and IndoBERT.

Table 3 shows that IndoBERT outperformed the TF-IDF + Logistic Regression baseline in accuracy, precision macro, and F1 macro. IndoBERT achieved an accuracy of 0.8920, while the baseline model achieved 0.8092. The F1 macro also improved from 0.5666 to 0.6347.

To further validate whether the performance improvement was statistically significant, McNemar's test was conducted on the same test set. The results are shown in Table 4 and Figure 3.

Table 4. McNemar test results for model comparison.

Item	Value
Both models correct	9,024
Baseline correct, IndoBERT wrong	286
Baseline wrong, IndoBERT correct	1,239
Both models wrong	956
Discordant pairs	1,525
McNemar chi-square statistic	594.30
Exact McNemar p-value	< 0.001
Interpretation	Statistically significant

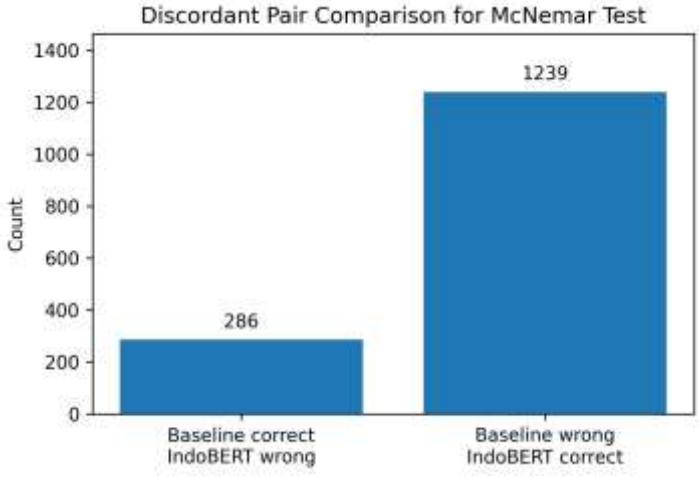

Figure 3. Discordant pair comparison for McNemar test

The McNemar test showed that both models correctly classified 9,024 instances, while both models incorrectly classified 956 instances. More importantly, IndoBERT correctly classified 1,239 instances that were incorrectly classified by the baseline model, whereas the baseline model correctly classified only 286 instances that were incorrectly classified by IndoBERT. The McNemar test produced a chi-square statistic of 594.30 with an exact p-value below 0.001. Therefore, the performance difference between the two models was statistically significant.

The confusion matrix of IndoBERT is presented in Figure 4, and the class-level performance is shown in Table 5.

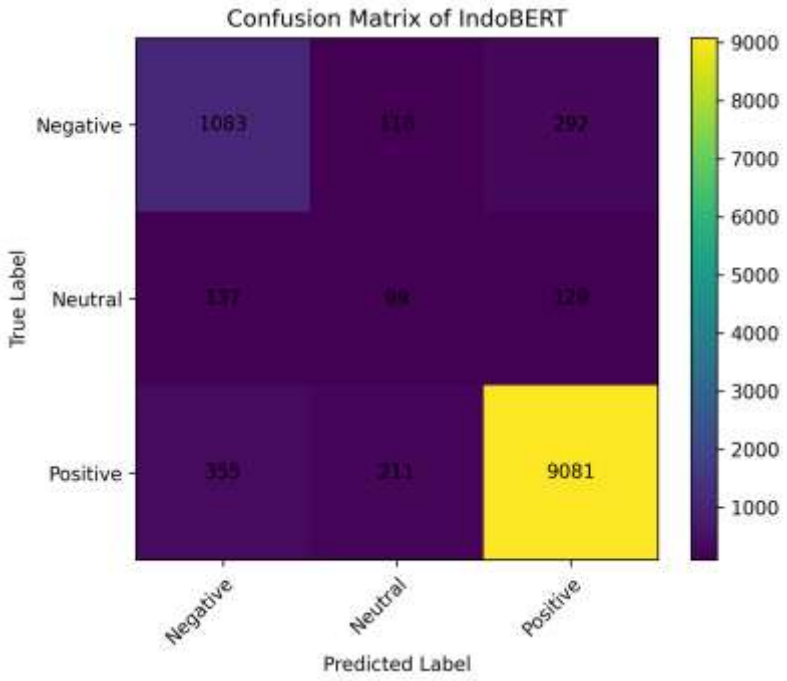

Figure 4. Confusion matrix of IndoBERT

Table 5. Class-level performance of IndoBERT

Class	Support	Precision	Recall	F1-score
Negative	1,493	0.6876	0.7254	0.7060
Neutral	365	0.2313	0.2712	0.2497
Positive	9,647	0.9557	0.9413	0.9485

IndoBERT achieved the highest performance in the positive class, with a precision of 0.9557, recall of 0.9413, and F1-score of 0.9485. The negative class achieved moderate performance with an F1-score of 0.7060, while the neutral class obtained the lowest F1-score of 0.2497. This result indicates that neutral reviews were more difficult to classify because they often contained ambiguous, short, or mixed expressions.

After sentiment classification, BERTopic was used to discover latent topics in the user reviews. The model generated 41 topic groups, including one outlier topic. The ten largest non-outlier topics are presented in Table 6, while their sentiment proportions are visualized in Figure 5.

Table 6. Top 10 BERTopic topics and sentiment proportions

Topic	Count	Main Keywords	Dominant	Negative (%)	Neutral (%)	Positive (%)
0	7,989	mantap, suka, keren, ya, mau, kok	Positive	27.86	4.46	67.68
1	3,779	belajar, membantu, sangat, pembelajaran, mudah	Positive	1.85	0.48	97.67
2	3,741	video, videonya, animasi, vidio, belajar	Positive	14.57	13.10	72.33
3	3,130	sangat, belajar, membantu, bagus, anak	Positive	1.09	0.64	98.27
4	2,731	bagus, bermanfaat, sangat bagus, good	Positive	1.76	0.66	97.58
5	2,702	pelatihan, skill academy, prakerja	Positive	0.44	0.22	99.33

6	1,929	ya, kenapa, mantap, tapi, mau	Positive	28.25	6.53	65.22
7	1,665	belajar, sangat membantu, aplikasinya	Positive	0.42	0.60	98.98
8	1,598	mudah, dipahami, mudah dipahami	Positive	1.06	0.38	98.56
9	1,468	bintang, kasih bintang, lima	Positive	14.03	7.90	78.07

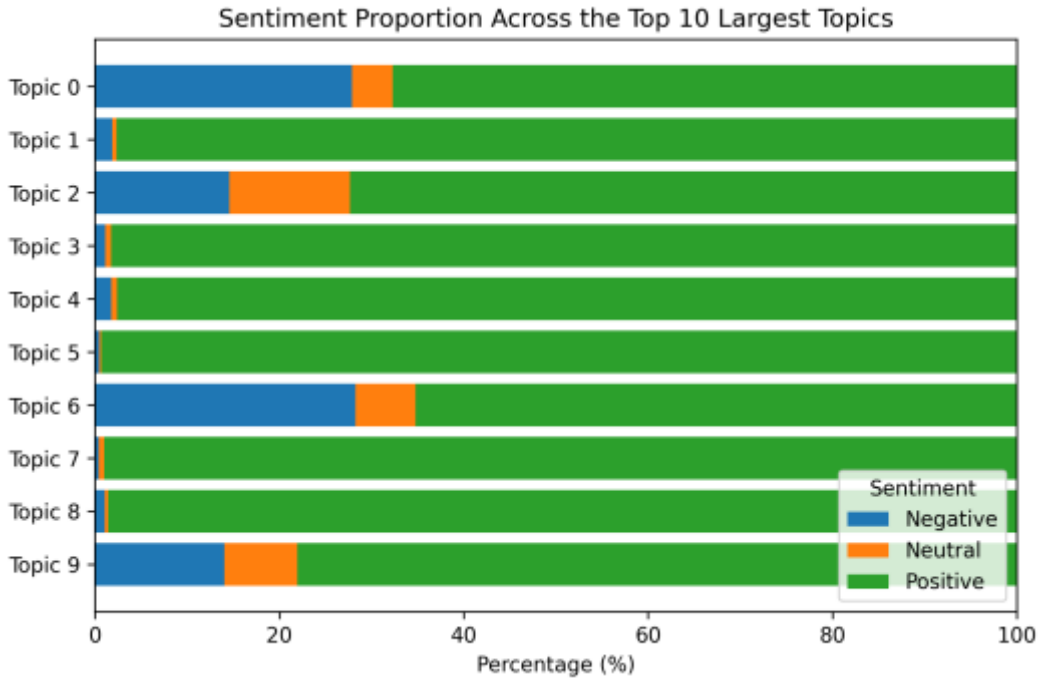

Figure 5.Sentiment proportion across the top 10 largest topics

Table 6 shows that most of the largest topics were dominated by positive sentiment. Topic 1, which was related to learning assistance and ease of learning, contained 97.67% positive sentiment. Topic 3 had 98.27% positive sentiment, and Topic 5, which was related to Skill Academy and Prakerja training, showed the highest positive sentiment proportion among the top topics with 99.33%.

However, several topics also contained relatively high negative sentiment proportions. Topic 0 had 27.86% negative sentiment, while Topic 6 had 28.25% negative sentiment. Topic 2, which was related to videos and animation-based learning, had 14.57% negative sentiment and 13.10% neutral sentiment, suggesting that video-based learning was generally appreciated but still generated some concerns. Smaller topics not shown in Table 6 also revealed dissatisfaction related to payment, application updates, advertisements, and account-related issues.

3.2. ANALYSIS

The results indicate that IndoBERT performed significantly better than the TF-IDF + Logistic Regression baseline for sentiment classification of Indonesian Ruangguru user reviews. The improvement was observed not only in descriptive evaluation metrics, where IndoBERT achieved higher accuracy and F1 macro, but also in the McNemar test result. The test showed a statistically significant difference between the two classifiers, with IndoBERT correctly classifying substantially more instances that were misclassified by the baseline model. This finding strengthens the argument that contextual representation contributes meaningfully to sentiment classification performance in Indonesian user-generated reviews.

This performance improvement can be explained by the characteristics of user reviews. Ruangguru reviews often contain informal Indonesian expressions, short comments, spelling variations, and implicit emotional meanings. TF-IDF mainly represents text through word frequency patterns, while IndoBERT captures contextual relationships among words. As a result, IndoBERT is more capable of identifying sentiment patterns that depend on context rather than surface-level word occurrence alone.

Nevertheless, the model still faced challenges in classifying neutral sentiment. The neutral class had the lowest precision, recall, and F1-score. This may be caused by class imbalance and the ambiguous nature of neutral reviews. Some neutral reviews may contain questions, mixed opinions, or short comments, making them difficult to distinguish from negative or positive reviews.

The overall sentiment distribution suggests that user perception toward Ruangguru was generally positive. Most users appreciated the application because it supported learning activities, helped them understand materials, and provided useful video-based explanations. This result is consistent with the topic modeling output, where many dominant topics were associated with learning support, ease of understanding, helpful materials, and application usefulness.

The BERTopic results provide deeper insights than sentiment classification alone. Positive sentiment was strongly associated with learning assistance, material comprehension, video learning, and Skill Academy training. Negative sentiment was concentrated in more specific issues such as payment, application updates, bugs, advertisements, and account-related problems. These findings indicate that topic-sentiment mapping can help developers identify both strengths and priority areas for application improvement.

The presence of a large outlier topic also requires attention. In BERTopic, outliers represent documents that do not belong to dense semantic clusters. A large outlier group may occur when many reviews are short, generic, repetitive, or semantically broad. Therefore, future studies may improve BERTopic performance by applying more advanced text normalization, phrase extraction, stopword enrichment, or hyperparameter tuning.

4. CONCLUTION

This study analyzed user reviews of the Ruangguru application using IndoBERT for sentiment classification and BERTopic for topic modeling. The dataset originally contained 100,000 Kaggle reviews, and 76,699 valid reviews were used after cleaning and preprocessing. The sentiment distribution showed that positive reviews dominated the dataset, indicating that most users expressed favorable perceptions of Ruangguru.

The experimental results showed that IndoBERT outperformed the TF-IDF + Logistic Regression baseline. IndoBERT achieved an accuracy of 0.8920 and an F1 macro score of 0.6347, while the baseline model achieved an accuracy of 0.8092 and an F1 macro score of 0.5666. The McNemar test further confirmed that this performance difference was statistically significant, with a chi-square statistic of 594.30 and an exact p-value below 0.001. These results indicate that contextual language representation provides a significant advantage over conventional word-frequency-based features for classifying Indonesian Ruangguru user reviews.

The topic modeling results showed that BERTopic generated 41 topic groups, including one outlier topic. Most dominant topics were associated with positive user experiences, such as learning support, material comprehension, video-based learning, and application usefulness. However, several topics showed strong negative sentiment, particularly those related to payment, application updates, technical problems, advertisements, and account issues.

The findings demonstrate that combining IndoBERT and BERTopic provides a more comprehensive understanding of user perceptions than sentiment analysis alone. This approach can help educational application developers identify not only whether users are satisfied or dissatisfied, but also what specific aspects generate those sentiments.

5. FUTURE WORKS

Future studies can improve this research in several directions. First, class imbalance should be addressed using techniques such as class weighting, oversampling, data augmentation, or focal loss to improve the classification of minority classes, especially neutral sentiment. Second, future research can compare IndoBERT

with other Indonesian Transformer models, such as IndoBERTweet, IndoRoBERTa, or multilingual BERT, to obtain a more comprehensive model comparison.

Third, BERTopic performance can be improved through hyperparameter tuning, embedding model comparison, and topic coherence evaluation. Fourth, future studies can conduct aspect-based sentiment analysis to identify more specific aspects, such as pricing, video quality, learning content, application performance, and customer service. Fifth, future research can compare Ruangguru with other educational applications to evaluate user perception across different digital learning platforms.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the availability of the Ruangguru user review dataset from Kaggle, which made this study possible. The authors also acknowledge the open-source NLP libraries and pre-trained language models used in this research, including IndoBERT and BERTopic. No specific funding was received for this study.

REFERENCES

Abel, N., Afifah, A., & Ayunda, T. (2026). PENERAPAN INDOBERT DAN BERTOPIC DALAM ABSA UNTUK EVALUASI PENDAHULUAN. *RABIT : Jurnal Teknologi Dan Sistem Informasi Univrab*, *11*(1), 847–868. https://doi.org/10.36341/rabit.v11i1.7143

Adam, A. Z. R., & Setiawan, E. B. (2023). Social Media Sentiment Analysis Using Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU). *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *9*(1), 119–131. https://doi.org/10.26555/jiteki.v9i1.25813

Adam, N. P., Gernowo, R., & Surarso, B. (2026). Integration of BERTopic and IndoBERTweet for Aspect-Based Sentiment Analysis (ABSA) on Short Text Data : A Case Study of Responses to Government Policies in 2025. *Journal of Economics, Technology and Business*, *5*(1), 37–53.

Anugrah, A. F., & Agatha, R. D. (2025). Utilizing IndoBERT and BERTopic to Explore Public Opinion on BPS Instagram Posts. *Journal of Applied Informatics and Computing*, *9*(5), 2836–2844. https://doi.org/10.30871/jaic.v9i5.10327

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 8875–8898. https://doi.org/10.18653/v1/2021.emnlp-main.699

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*(May), 1–16. https://doi.org/10.3389/fsoc.2022.886498

Geni, L., Yulianti, E., & Sensuse, D. I. (2023). Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *9*(3), 746–757. https://doi.org/10.26555/jiteki.v9i3.26490

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://doi.org/10.48550/arXiv.2203.05794

Hanny, D., & Resch, B. (2024). Clustering-Based Joint Topic-Sentiment Modeling of Social Media Data: A Neural Networks Approach. *Information (Switzerland)*, *15*(4), 1–20. https://doi.org/10.3390/info15040200

Jayadianti, H., Kaswidjanti, W., Utomo, A. T., Saifullah, S., Dwiyanto, F. A., & Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. *ILKOM Jurnal Ilmiah*, *14*(3), 348–354. https://doi.org/10.33096/ilkom.v14i3.1505.348-354

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, *6*(February), 100059. https://doi.org/10.1016/j.nlp.2024.100059

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweeT : A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *Proceeeding of the 2021 Converence on Empirical Methods in Natural Language Processing*, 10660–10668.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Li, C., & Hu, X. (2025). Medical artificial intelligence in scholarly and public perspective: BERTopic-based analysis of topic-sentiment collaborative mining. *Data Science and Informetrics*, *5*(1), 33–42. https://doi.org/10.1016/j.dsim.2025.05.001

Mayzaroh, M. A., Ningsih, D. F., Destriani, N., & Manullang, M. C. T. (2026). *Benchmarking PyCaret AutoML Against IndoBERT Fine-Tuning for Sentiment Analysis on Indonesian IKN Twitter Data*. (1), 1–7. http://arxiv.org/abs/2604.25392

Pinilih, N. W., Dewi, I. N., & Alzami, F. (2026). *An Integrated Topic – Sentiment Analysis of User Reviews in Vidio*

Application Using BERTopic and IndoRoBERTa. *10*(3), 2685–2698.

Purwanto, D. D. (2026). *Empirical Evaluation of IndoBERT and LSTM for Sentiment Analysis of Tourism Reviews : A Data-Driven Study on Kenjeran Park*. *7*(1), 463–474.

Ramamoorthy, T., Kulothungan, V., & Mappillairaju, B. (2024). Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1329185

Saputra, M. A. A., Alamsyah, A., Ramadhani, D. P., Siadari, T. S., & Fakhrurroja, H. (2026). *IndoBERT-Sentiment: Context-Conditioned Sentiment Classification for Indonesian Text*. 1–8. http://arxiv.org/abs/2604.07057

Sihombing, W. M., & Widiyaningtyas, T. (2025). Sentiment Analysis and Topic Modelling Using IndoBERTweet and BERTopic for Public Health Issues. *Indonesian Journal of Innovation Studies*, *26*(4). https://doi.org/10.21070/ijins.v26i4.1833

Wafda, A., Fudholi, D. H., & Nugraha, J. (2025). Aspect-Based Sentiment Analysis on Twitter Tweets About the Merdeka Curriculum Using Indobert. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(3), 586–599. https://doi.org/10.33480/jitk.v10i3.5692