Optimization and Evaluation of IndoBERT, BiLSTM–BiGRU, and Hybrid XGBoost for Predicting Complaint Subcategories in the E-Layanan System of Universitas Negeri Surabaya

Intan Alpiana[a], Wiyli Yustanti[b], Yuni Yamasari[c]

[a,b,c] *Department of Informatics, Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia*
email: [a]intanalpiana@unesa.ac.id, [b]wiyliyustanti@unesa.ac.id, [c]yuniyamasari@unesa.ac.id

Abstrak

Peningkatan kualitas layanan digital menjadi salah satu prioritas dalam mendukung kegiatan akademik di perguruan tinggi. Universitas Negeri Surabaya telah mengimplementasikan sistem E-Layanan sebagai sarana penyampaian keluhan civitas akademika terkait penggunaan sistem informasi dan jaringan. Namun, proses klasifikasi subkategori keluhan masih dilakukan secara manual sehingga menyebabkan keterlambatan penanganan dan meningkatnya beban admin. Penelitian ini bertujuan mengoptimalkan model IndoBERT untuk memprediksi subkategori keluhan secara otomatis berdasarkan teks deskripsi pengguna. Tiga pendekatan dievaluasi, yaitu IndoBERT sebagai baseline, IndoBERT + BiLSTM + BiGRU sebagai pendekatan sekuensial dua arah, dan Hybrid IndoBERT + XGBoost sebagai kombinasi embedding IndoBERT dengan algoritma klasifikasi XGBoost. Lebih dari 8000 data keluhan periode 2017–2024 digunakan dalam proses pelatihan dan evaluasi. Hasil menunjukkan bahwa IndoBERT menghasilkan akurasi 83%, sedangkan dua pendekatan optimalisasi mencapai akurasi 95%. Temuan ini membuktikan bahwa kombinasi IndoBERT dengan teknik sekuensial dan hybrid meningkatkan akurasi prediksi serta mempercepat respons keluhan pada sistem E-Layanan UNESA.

Kata Kunci: IndoBERT; Klasifikasi Teks; Subkategori Keluhan.

Abstract

Improving the quality of digital services is essential to support academic activities in universities. Universitas Negeri Surabaya has implemented the E-Layanan system to manage complaints related to information systems and networks. However, the classification of complaint subcategories is still conducted manually, causing response delays and increasing administrative workload. This study aims to optimize IndoBERT to automatically predict complaint subcategories based on textual descriptions. Three approaches were evaluated: IndoBERT as the baseline, IndoBERT + BiLSTM + BiGRU as a bidirectional sequential approach, and Hybrid IndoBERT + XGBoost as a combination of IndoBERT embedding and XGBoost classification. More than 8000 complaint records from 2017 to 2024 were used in the evaluation. The results show that IndoBERT achieved an accuracy of 83%, while both optimized approaches reached 95%. These findings indicate that combining IndoBERT with sequential and hybrid techniques improves prediction accuracy and enables faster and more accurate responses to user complaints

Keywords: *IndoBERT; Text Classification; Complaint Subcategory.*

INTRODUCTION

The development of digital technology has pushed universities to improve the quality of academic services based on information systems. Universitas Negeri Surabaya (Unesa) is one of the institutions that has implemented E-Layanan, a digital platform that allows the academic community to submit complaints related to information systems, networks, and other academic services. The presence of this platform is expected to accelerate the communication process and problem handling. However, the classification of complaint subcategories on E-Layanan is still done manually by the admin, which often leads to delays in handling, inconsistency in classification, and an increased workload. This condition highlights the importance of applying Natural Language Processing (NLP) to automate the process of classifying complaint texts (Agarwal, Kumar, & Ganguly, 2023).

In previous research, various text classification approaches have been used, ranging from classical machine learning such as Naïve Bayes, Support Vector Machine (SVM), and Random Forest, to more complex deep learning methods. One of the widely used models is BERT (Bidirectional Encoder Representations from Transformers), which has been proven to excel in understanding linguistic context (Devlin et al., 2019). The adaptation version for the Indonesian language, namely IndoBERT, shows significant results on various NLP tasks, including sentiment analysis, text classification, and entity extraction (Wilie et al., 2020; Jayadianti et al., 2022; Asri et al., 2025; Sari & Utomo, 2023).

Nevertheless, there are significant challenges in implementing NLP for the Indonesian language. First, the structure of the Indonesian language differs from that of English, such as the use of affixes, reduplication, and a very broad vocabulary variation (Rahma & Suadaa, 2023). Second, complaint data in E-Layanans tends to be imbalanced, as there are subcategories with a large amount of data, while other subcategories only have a small amount of data. This imbalance can reduce the performance of classification models, as models tend to be biased towards the majority class (He & Garcia, 2009; Wei & Zou, 2019; Longpre, Lu, & Tan, 2020).

To address this issue, this research performs data balancing before the model training process. Several subcategories with very limited data are combined into an 'OTHER' category, while the distribution between classes is balanced using oversampling techniques and text augmentation such as back-translation and synonymization (Rahma & Suadaa, 2023). This strategy aims to increase the amount of data in the minority classes to prevent the model from being biased towards the majority class, while still maintaining the semantic diversity of complaint data. Thus, the dataset used becomes more proportional and representative of the variety of complaints found in Unesa's E-Layanan.

In addition, complaint data in the form of raw text usually contains noise such as typos, non-standard abbreviations, the use of emoticons, and excessive spacing. Therefore, a text preprocessing stage is necessary before the data is used for model training. This stage includes word normalization, removal of punctuation, elimination of double spaces, and standardization of word forms. With good preprocessing, the quality of text representation can be improved, thus supporting the model's performance in understanding the context of user complaints (Yustanti et al., 2024; Yamasari et al., 2024).

In the context of model development, recent research has largely combined IndoBERT with sequential methods such as BiLSTM (Bidirectional Long Short-Term Memory) and BiGRU (Bidirectional Gated Recurrent Unit). Both methods are capable of capturing bidirectional contextual relationships in text, thereby enriching the feature representations produced by IndoBERT (Liu et al., 2016; Cho et al., 2014; Rahman et al., 2024). Additionally, the Hybrid IndoBERT + XGBoost approach has also garnered attention, as it combines the power of semantic representation from transformers with the efficient classification capabilities of the XGBoost algorithm, which is known to excel in handling high-dimensional data (Talaat, 2023; Akanbi et al., 2025).

Based on the problem, this research focuses on the optimization and evaluation of three methods, namely: (1) IndoBERT as a baseline, (2) IndoBERT + BiLSTM + BiGRU as a bidirectional sequential approach, and (3) Hybrid IndoBERT + XGBoost. More than 8,000 complaint data collected from Unesa's E-Layanans during the period of 2017–2024 are used as training and evaluation material.

The main objective of this research is to compare the performance of the three methods in automatically predicting complaint subcategories, as well as to identify the most effective method to support the improvement of digital services at Unesa. Thus, this research not only contributes to the development of NLP technology for the Indonesian language but also provides practical solutions to enhance the responsiveness of digital-based academic service systems.

METHOD

This research design will compare several text classification modeling methods to obtain the best model in predicting the subcategories of student complaints. The research population consists of all student complaint data recorded in the e-service of Universitas Negeri Surabaya with a total of over 8000 complaint data. The research sample is taken in full (total sampling), where all data is used as research objects because it can still be processed by natural language processing models. From the total data, a division is made into training data (80%) and test data (20%) using a stratified split technique to ensure the distribution of categories remains proportional.

The data collection technique was conducted through documentation from the e-service system. Each entry contains text complaints regarding the use of the information system and network along with subcategory labels. The modeling algorithm in this study is IndoBERT (transformer-based classification) (Devlin et al., 2019; Wilie et al., 2020), BiLSTM–BiGRU (RNN-based deep learning) (Rahman et al., 2024) and Hybrid IndoBERT–XGBoost (a combination of transformer & ensemble learning) (Talaat, 2023; Akanbi et al., 2025). Data analysis is conducted based on the stages of CRISP-DM, which are as follows:

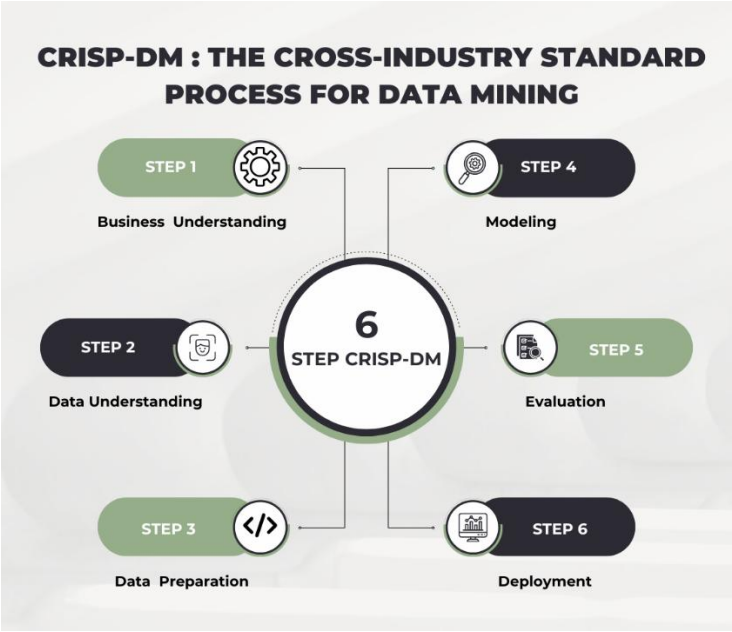


Fig 1. The flow of the CRISP-DM framework

The data analysis process in this research refers to the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which consists of six stages: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. This

framework is chosen because it is systematic, flexible, and aligns with the characteristics of data-driven research and computational experiments.

2.1 Business Understanding

This stage focuses on formulating research objectives, namely optimizing and evaluating three text classification approaches: IndoBERT, BiLSTM–BiGRU, and Hybrid XGBoost in predicting complaint subcategories in the e-services of Universitas Negeri Surabaya. The main issue raised is the large number of complaint subcategories with an imbalanced data distribution, thus requiring balancing strategies and the selection of the appropriate classification method (Rahma & Suadaa, 2023).

2.2 Data Understanding

The data used consists of over 8000 entries of complaints from students and lecturers regarding information system and network services. In this phase, initial data exploration, class distribution analysis, and identification of issues such as duplication, missing data, and class dominance are conducted (Jayadianti et al., 2022; Asri et al., 2025).

2.3 Data Preparation

This stage is the most crucial process to ensure that the data is ready to be used in the model. The steps taken include:

a) Text normalization: removal of excess spaces, numbers, irrelevant punctuation, and standardization of lowercase letters (Yustanti et al., 2024).
b) Balancing data: subcategories with less than 80 data points are merged into the OTHER category, followed by augmentation techniques (e.g., back-translation) to achieve a more balanced class distribution (Wei & Zou, 2019; Longpre et al., 2020; Rahma & Suadaa, 2023).
c) Label encoding: each complaint subcategory is encoded into a numeric label.
d) Split data: the dataset is divided into training data (80%) and test data (20%).

2.4 Modeling

This stage is the core of the research, where the implementation of three different methods is conducted. The flow is adjusted according to the implementation code used, which is:

a. IndoBERT: Text tokenization using the IndoBERT tokenizer, text representation with the [CLS] token as a sentence vector, addition of a linear classification layer, and training the model with the AdamW optimizer and cross-entropy loss (Devlin et al., 2019; Wilie et al., 2020).
b. BiLSTM–BiGRU: Feature extraction from text through an embedding layer, sequential processing using BiLSTM and BiGRU layers to capture bidirectional context (Liu et al., 2016; Cho et al., 2014), and a fully connected layer with a softmax activation function for category prediction.
c. Hybrid XGBoost: Text representation is transformed into TF-IDF vectors, and the XGBoost model is used for classification, with parameters optimized through grid search to improve accuracy (Talaat, 2023).

2.5 Evaluation

The evaluation is conducted objectively using accuracy metrics to measure the percentage of correct predictions. Precision, Recall, and F1-Score are used to evaluate the performance per class. A Confusion Matrix is utilized to view the distribution of classification errors. Top-3 Accuracy specifically for the IndoBERT model to measure the model's ability to provide relevant alternative predictions.The evaluation results are presented in the form of tables and graphs making them easier to analyze (Rahman et al., 2024; Sari & Utomo, 2023).

2.6 Deployment

The final stage involves interpreting the results and compiling recommendations for the best model that can be adopted by the UNESA e-service managers. The model with the highest performance

can be used as a prototype for an automatic classification system to assist in handling complaints (Akanbi et al., 2025; Agarwal et al., 2023).

RESULTS AND DISCUSSION

3.1 Result

This research uses a dataset of user complaints from the E-Service system of Universitas Negeri Surabaya (UNESA), which consists of more than 8000 complaint data entries. Each entry represents a complaint categorized into dozens of subcategories, such as 'Login Issues', 'Data Errors', 'Slow Response', and so on. However, the initial distribution among subcategories is unbalanced: there are subcategories with less than 20 data entries, while several major categories have hundreds. To maintain the validity of the model and avoid bias towards the majority class, several preprocessing steps were taken as follows:

a. Text normalization: the entire text is converted to lowercase (case folding), removing punctuation, double spaces, and non-alphabetic characters (Yustanti et al., 2024; Yamasari et al., 2024).
b. Removal of duplicates and noise: duplicate data and very short/insignificant entries are removed (Jayadianti et al., 2022; Asri et al., 2025).
c. Merging minority categories: all subcategories with less than 80 data entries are merged into a new label called OTHERS, to maintain balance in the amount of data among classes (Rahma & Suadaa, 2023).
d. Splitting training and testing data: the dataset is divided into 80% training data and 20% testing data using a stratified split to ensure class distribution is preserved in both subsets (Sari & Utomo, 2023). After this process, a cleaner and more balanced dataset was obtained, ready to be used for training and evaluating classification models (Longpre et al., 2020).

This research compares three approaches: the first is IndoBERT (baseline), a BERT model specifically designed for Indonesian without additional architecture or optimization methods (Wilie et al., 2020). The second is IndoBERT + BiLSTM + BiGRU (Ensemble), which is a combination of three models; IndoBERT is used to generate contextual embeddings, BiLSTM captures bidirectional sequential dependencies, and BiGRU enhances long-term memorization, which are then combined in an ensemble using voting or averaging techniques (Rahman et al., 2024). Lastly, the third approach is IndoBERT + XGBoost, where the IndoBERT model is used to generate embeddings (vector representations), and then the results are classified using the XGBoost algorithm known to be effective in handling gradient boosting-based classification (Talaat, 2023). Below is a graph of the accuracy results from the three approaches.

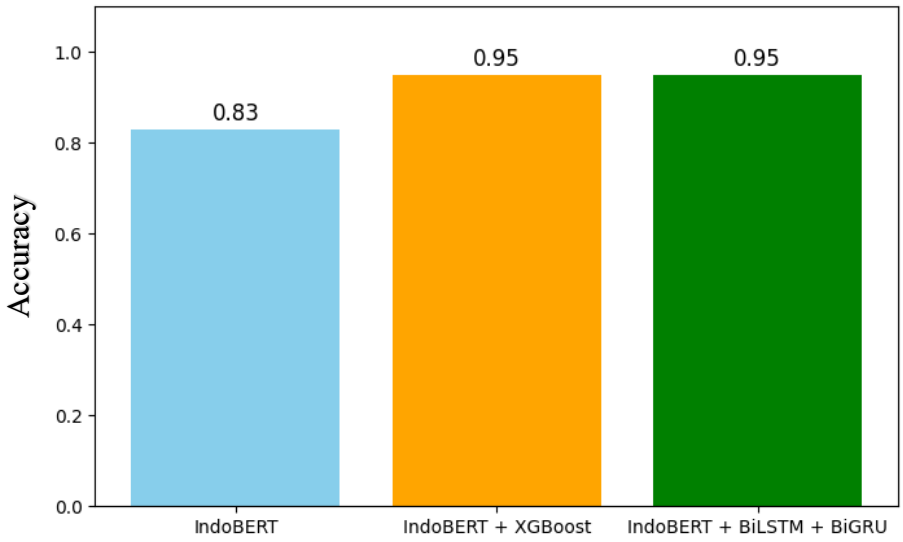


Fig 1. Comparison of model accuracy

This result shows that both approaches optimizing the IndoBERT representation (either through XGBoost or RNN ensemble) experienced a significant increase of +12% compared to the IndoBERT baseline.

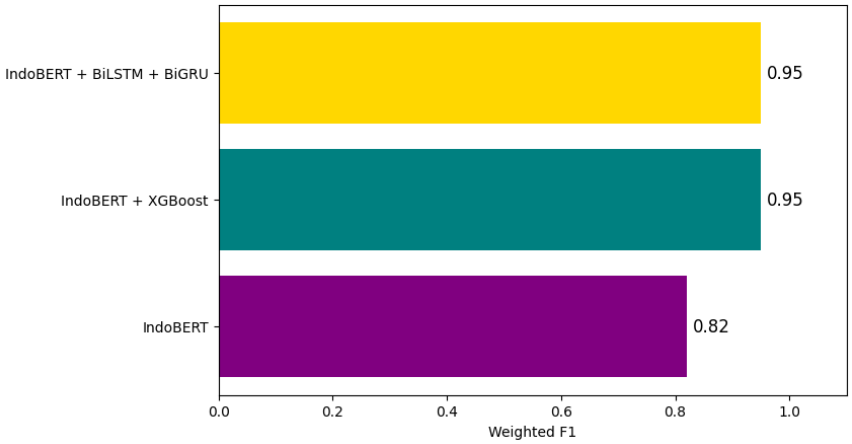


Fig 2. Comparison of weighted F1-score

Table 1. Comparison of the results of the three methods

Model	**Total Accuracy**	**Macro Avg Precision**	**Macro Avg Recall**	**Macro Avg F1-score**	**Weighted Avg F1-score**
IndoBERT	83%	0.86	0.87	0.87	0.82
IndoBERT+XGBoost	95%	0.95	0.95	0.95	0.95
IndoBERT + BiLSTM + BiGRU (Ensemble)	95%	0.95	0.95	0.95	0.95

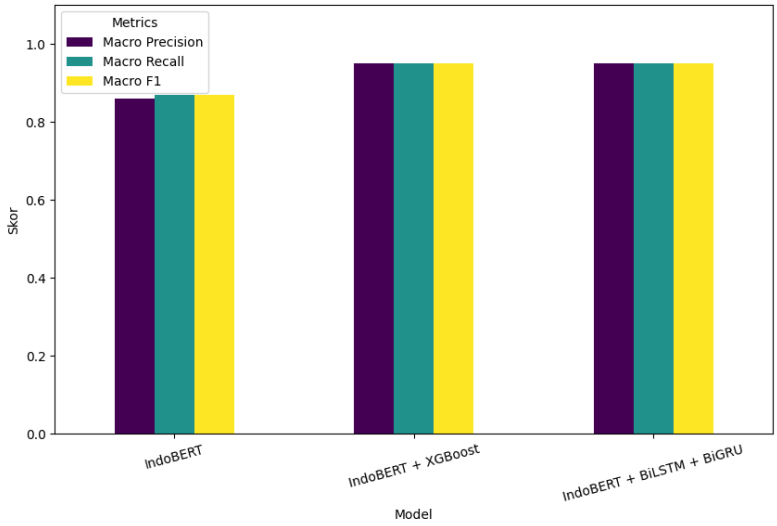


Fig 3. Comparison of macro average metrics

Error analysis was conducted to understand the weaknesses of each model. Some key findings are that IndoBERT makes many misclassifications in minority categories such as "OTHERS" or semantically similar subcategories, for example between "Wrong Input" and "Wrong Data". IndoBERT + XGBoost shows a more even distribution of errors; most errors come from classes that are very semantically similar or indeed ambiguous. IndoBERT + BiLSTM + BiGRU (Ensemble) has the fewest errors and tends to be more stable in classifying both minority and majority classes, thanks to the combination of contextual embedding power and bidirectional sequential processing.

3.2 Discussion

a. Baseline Evaluation: IndoBERT

The IndoBERT model achieves an accuracy of 83%, which can be considered quite good for a baseline. IndoBERT is a variant of BERT that has been specifically trained on a corpus of the Indonesian language, enabling it to understand the syntactic and semantic structure of the local language well (Wilie et al., 2020). However, its main limitation lies in the absence of a specific mechanism to handle class distribution imbalance, the lack of additional layers that explicitly model word order in a context of length, and sensitivity to minority classes and very short or ambiguous complaint texts. This condition causes the model to be biased towards the majority class (Wilie et al., 2020).

b. Ensemble Approach: BERT + BiLSTM + BiGRU

The third approach uses an ensemble of BERT, BiLSTM, and BiGRU. The scheme is that IndoBERT generates contextual embeddings for each token (Wilie et al., 2020). These embeddings are then passed to two bidirectional RNN architectures: BiLSTM and BiGRU. The outputs of both are combined, either through concatenation or voting, before being forwarded to the final classification layer. The reason for using BiLSTM and BiGRU together is that BiLSTM excels at capturing long-term dependencies both forward and backward but is relatively computationally heavy, whereas BiGRU has a simpler structure with nearly equivalent performance but faster training time (Longpre et al., 2020). The combination of the two creates diversity in sequential representation, which is key to the success of ensemble systems. The advantages of this approach are that it is more capable of handling variations in both long and short sentence structures, is more stable with minority classes due to richer context representation, and can achieve 95% accuracy, similar to BERT + XGBoost, but with a stronger depth of context understanding. The drawbacks are longer training and inference times, as well as a greater need for computational resources (Rahma & Suadaa, 2023).

c. Optimization of IndoBERT Representation + XGBoost

The IndoBERT + XGBoost approach separates the representation and classification processes, where IndoBERT is used only as a feature extractor, producing embedding vectors that represent the complete context of sentences (Wilie et al., 2020). XGBoost is used to classify these vectors. XGBoost excels at handling high-dimensional data, performs regularization to avoid overfitting, and is efficient in computation. The advantage of this approach is that it is faster during the inference stage, as BERT is not used for backpropagation during classification, is more robust against noise and class distribution, as XGBoost can balance class weights through boosting, and this approach can achieve an accuracy of 95%, an increase of 12% from the baseline. These results are consistent with previous studies that show combining embeddings from transformer models with classic boosting algorithms can improve performance in domain-specific text classification (Wilie et al., 2020).

Table 2. Comparison of the results of the three methods

Aspect	**IndoBERT**	**IndoBERT + XGBoost**	**IndBERT + BiLSTM + BiGRU**
Accuracy	83%	95%	95%
Inference speed	Fast	Very fast	At the latest
Model complexity	Low	Medium	Height
Sequence context capture	Limited	Low	Very good
Suitable for real-time production	Yes	Very suitable	Not recommended (heavy)

The two best models (BERT + XGBoost and Ensemble RNN) have the same accuracy, but are suitable for different scenarios: IndoBERT + XGBoost is more suitable for practical implementation in E-Service systems because it is fast, light, and accurate. Ensemble RNN is more suitable for research experiments or deep analysis systems, as it is capable of capturing deeper context even though it is slower. This research confirms that the combination of transformer models with other classification algorithms (XGBoost, RNN) can provide significant improvements in the performance of text classification in the Indonesian language. This expands the literature on NLP in non-English languages, particularly Indonesian. With accuracy of up to 95%, the model can be integrated into E-Service systems to automatically categorize complaints, provide quick notifications to relevant units, shorten complaint handling time, and reduce the manual workload of operators/admins.

CONCLUSIONS

This research aims to evaluate and compare the performance of various transformer-based approaches and hybrid methods in classifying complaint subcategories in information systems and network services at Surabaya State University. The results show that the pure IndoBERT model can achieve a good accuracy of 83%, but when combined with the XGBoost algorithm and the ensemble architecture of BiLSTM and BiGRU, its performance significantly improves to achieve an accuracy of 95% with high consistency in precision, recall, and f1-score values. This finding proves that the integration of the IndoBERT-based model with advanced learning methods can produce more accurate and reliable classifications compared to single modeling. Thus, the research hypothesis regarding the effectiveness of the hybrid approach in enhancing classification performance is validated. This study also confirms that the main objective, which is to obtain an optimal classification model to support the analysis of users' complaints in information systems, has been successfully achieved. Going forward, further research could be directed towards testing real data scenarios on a larger scale, developing models with explainable AI techniques to improve interpretability, and applying models in automatic complaint systems to provide practical benefits for managing digital services in universities.

ACKNOWLEDGMENT

The author would like to express the deepest gratitude to Universitas Negeri Surabaya (UNESA) through the Faculty of Engineering for providing support, facilities, and opportunities to conduct this research. Thanks are also extended to the team at the Directorate of Information and Communication Technology at UNESA for providing important data and information to support the research process. Additionally, sincere appreciation is directed to the supervising lecturers and colleagues who have offered feedback, guidance, and motivation during the preparation of this article. Their support, suggestions, and assistance have been invaluable in producing more focused and comprehensive research.

REFERENCES

Agarwal, S., Kumar, A., & Ganguly, R. (2023). Investigating transformer-based models for automated e-governance in Indian Railway using Twitter. Social Network Analysis and Mining, 13(1), Article 1–15. https://doi.org/10.1007/s13278-023-01037-1

Akanbi, C., & Co-Authors. (2025). Evaluating the Performance of BERT, XGBoost, and Hybrid Models for Fake News Detection.

Asri, Y., Kuswardani, D., Suliyanti, W. N., & Manullang, Y. O. (2025). Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application. Indonesian Journal of Electrical Engineering and Computer Science, 38(1), 677–688. https://doi.org/10.11591/ijeecs.v38.i1.677-688

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

Longpre, S., Lu, Y., & Tan, Y. C. (2020). How effective is task-agnostic data augmentation for pretrained transformers? arXiv preprint arXiv:2003.02245. https://arxiv.org/abs/2003.02245

Rahma, I. A., & Suadaa, L. H. (2023). Penerapan text augmentation untuk mengatasi data yang tidak seimbang pada klasifikasi teks berbahasa Indonesia. Jurnal Teknologi dan Sistem Komputer, 11(2), 95–102. https://doi.org/10.14710/jtsiskom.2023.14149

Rahman, M. M., Shiplu, A. I., Watanobe, Y., & Alam, M. A. (2024). RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis. arXiv preprint, 2406.00367. Retrieved from https://arxiv.org/abs/2406.00367

Sari, N. P., & Utomo, E. (2023). Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7(6), 1234–1244. https://doi.org/10.29207/resti.v7i6.4721

Talaat, A. S. (2023). Sentiment analysis classification system using hybrid BERT models. Journal of Big Data, 10, Article 110. https://doi.org/10.1186/s40537-023-00781-w

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP IJCNLP) (pp. 6382–6388). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1670

W. Yustanti, A. W. Utami, G. S. Palupi and P. S. Nautika, "Probabilistic-based Text Clustering for Optimizing Mental Health Issues Extraction on Social Media," 2024 Seventh International Conference on Vocational Education and Electrical Engineering (ICVEE), Malang, Indonesia, 2024, pp. 70-75, doi: 10.1109/ICVEE63912.2024.10823692.

Yamasari, Y., Qoiriah, A., Rochmawati, N., Prapanca, A., Prihanto, A., Suartana, I. M., & Ahmad, T. (2024). Exploring the tree algorithms to generate the optimal detection system of students' stress levels. Indonesian Journal of Electrical Engineering and Computer Science, 36(1), 548–558. https://doi.org/10.11591/ijeecs.v36.i1.pp548-558