

A Comparative Study of Machine Learning Algorithms for Indonesian Hoax Detection on Social Media

Zeny Widianingsih¹, Aldi Algino², Septien Dwi Savandha³

STMIK IKMI Cirebon, Indonesia^{1,2}

Universidad Tecnológica Latinoamericana en Línea (UTEL), Mexico³

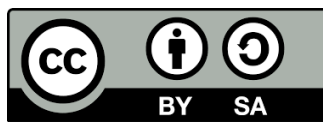
Corresponding Author: zenywidianingsih939@gmail.com, alginoaldi88@gmail.com,
dwisavandha9@gmail.com

Keywords:

hoax detection; text classification;
machine learning; TF-IDF; social
media

ABSTRACT

Social media platforms have accelerated the spread of misinformation (hoax) in Indonesia, undermining public trust and informed decision-making. This study aims to compare the performance of three text classification algorithms in detecting Indonesian-language hoax content on social media and to identify linguistic patterns distinguishing hoax from valid content. A publicly available corpus of 3,754 Indonesian tweets labeled as hoax or valid was used. Text was preprocessed through case folding, URL removal, and non-alphabetic character removal, then represented using TF-IDF with unigram and bigram features. Three classifiers, Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Machine, were trained and evaluated on an 80:20 stratified train-test split using accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). Results: Linear SVM achieved the highest accuracy (84.29%) and F1-score (0.9106), Logistic Regression achieved the highest precision (0.9118) and AUC (0.8156), while Naive Bayes achieved the highest recall (0.9967). Feature analysis revealed that debunking-related terms such as hoax, hoaks, mitos, and faktanya were strongly associated with the valid class, since such terms frequently appear in fact-checking tweets rather than in genuinely neutral news content. Conclusion: Model selection for hoax detection should be guided by the operational goal, favoring high-recall models such as Naive Bayes when the priority is minimizing missed hoaxes, and higher-precision models such as Logistic Regression when minimizing false accusations is prioritized; additionally, the reliance on debunking-style vocabulary as a valid-class signal highlights an important limitation of purely lexical models that future research should address.



This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

Copyright © 2026 by Author. Published by
Publikasi Indonesia

INTRODUCTION

Social media has become the main channel for the dissemination of information in Indonesia, but the ease of sharing content also accelerates the spread of misinformation or hoaxes. The Ministry of Communication and Information Technology notes that thousands of hoax content are spread every year in various sectors, including health, government policy, and politics, with significant spikes ahead of major events such as general elections (Nurhikam et al., 2023). The uncontrolled spread of hoaxes has the

potential to erode public trust in institutions, trigger social polarization, and influence public decision-making based on misinformation.

Various informatics research has developed an Indonesian-language hoax detection system using machine learning and natural language processing approaches. Prasetijo et al. (2023) developed a hoax detection system on Indonesian news sites using SVM and Stochastic Gradient Descent, being one of the initial references in this field. Setiawan et al. (2021) applied the Naive Bayes algorithm to detect the validity of news on social media Twitter, while Agustina et al. (2022) and Elisabeth et al. (2023) respectively applied Naive Bayes and SVM to detect fake news on certain topics, such as rising fuel prices. Deep learning-based approaches have also been explored: Yusuf and Suyanto (2022) used Long Short-Term Memory (LSTM) for the detection of hoax texts in Indonesian, while Azizah et al. (2023) compared the BERT, ALBERT, and RoBERTa transformer models, finding that ALBERT provided the highest accuracy (87.6%) compared to other transformer models.

In a more specific context, Nurhikam et al. (2023) applied the Random Forest algorithm to detect fake news related to the 2024 election, and Fernandes and Shita (2024) compared SVM and Random Forest to detect hoax news in a case study of a specific company. Ariatmanto and Rifai (2024) highlighted the importance of feature extraction techniques in improving Random Forest's accuracy for fake news detection, while Najwan et al. (2024) also applied Random Forest for fake news classification with an emphasis on comparing the performance of various feature extraction techniques. A recent study by Tobing et al. (2025) compared IndoBERT and mBERT for the detection of Indonesian-language political hoax news, while Nabil et al. (2026) applied a combination of TF-IDF and SVM for hoax classification. Beyond the context of direct hoax detection, Al Fachrozi and Tania (2025) and Fahrur et al. (2025) provide additional evidence that systematic comparisons between Naive Bayes, SVM, and decision tree algorithms remain relevant for various Indonesian text classification tasks, including sentiment analysis.

Although there is considerable literature on hoax detection in Indonesian, most studies focus on reporting a single accuracy metric without explicitly addressing the trade-off between precision and recall, even though these two measures have very different operational implications in the context of content moderation: missing hoaxes (low recall) risks allowing misinformation to continue to spread, while flagging valid content as hoaxes (low precision) risks silencing correct information (Ridho & Yulianti, 2024). In addition, there have not been many studies that have systematically analyzed the linguistic patterns underlying model classification decisions, including the potential that learning models recognize debunking or debunking writing styles, rather than solely the factual truth of the content (Mridha et al., 2021). Therefore, this study aims to (1) compare the performance of three text classification algorithms, namely Naive Bayes, Logistic Regression, and Linear SVM, in detecting Indonesian-language hoaxes on social media data with an explicit emphasis on precision-recall trade-offs, and (2) identify the terms that contribute the most to classification decisions to reveal potential linguistic bias in the model (Bashaddadh et al., 2025).

Despite the growing body of research on Indonesian-language hoax detection, several important gaps remain in the existing literature. Previous studies have primarily focused on developing and evaluating machine learning or deep learning models based on classification accuracy, including Support Vector Machine, Naive Bayes, Random Forest, and transformer-based approaches (Ariatmanto & Rifai, 2024; Azizah et al., 2023;

Bagus et al., 2023; Tobing et al., 2025). However, the operational implications of differences in precision and recall have received comparatively limited attention, particularly regarding the consequences of false-positive and false-negative predictions in social media content moderation. Furthermore, relatively few studies have examined how specific linguistic features influence classification decisions and whether machine learning models distinguish misinformation based on factual characteristics or merely recognize recurring lexical patterns. This limitation is particularly relevant to datasets containing fact-checking and debunking content, where terms associated with misinformation may appear frequently in legitimate posts. Consequently, a more comprehensive comparative evaluation is needed to examine both classification performance and the linguistic characteristics underlying model predictions, particularly in imbalanced Indonesian-language social media datasets.

The novelty of this study lies in integrating comparative machine learning performance evaluation with an examination of lexical patterns associated with hoax classification in Indonesian-language social media content. Rather than focusing exclusively on classification accuracy, this research systematically compares Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Machine using five evaluation metrics: accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). This approach provides a more comprehensive understanding of the trade-offs between detecting misinformation and minimizing the misclassification of legitimate content. Furthermore, the study examines the feature coefficients of Logistic Regression to identify linguistic patterns associated with hoax and valid classifications. Particular attention is given to the potential influence of debunking-related vocabulary, which may cause lexical classification models to recognize writing styles rather than factual accuracy. By combining performance evaluation with feature-level interpretation, this study provides additional empirical insight into the methodological limitations and practical considerations of machine learning-based hoax detection systems.

This study aims to comparatively evaluate the performance of three supervised machine learning algorithms, namely Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Machine, in detecting Indonesian-language hoax content on social media. Specifically, the research seeks to determine the comparative performance of these algorithms based on accuracy, precision, recall, F1-score, and AUC; examine the trade-offs between precision and recall in an imbalanced text classification dataset; identify the linguistic features most strongly associated with hoax and valid content through Logistic Regression coefficient analysis; and investigate potential lexical biases arising from the presence of debunking-related vocabulary in legitimate social media posts. Through these objectives, the study seeks to provide an empirical basis for selecting appropriate classification models according to different operational requirements in automated misinformation detection.

This research is expected to provide both theoretical and practical benefits for the development of automated misinformation detection systems. From a theoretical perspective, the study contributes to the literature on natural language processing and supervised machine learning by providing comparative empirical evidence concerning the performance characteristics of three widely used classification algorithms in Indonesian-language text classification. It also contributes to a better understanding of the relationship between lexical feature representation, class imbalance, and classification outcomes. From a methodological perspective, the research demonstrates the importance

of evaluating machine learning models using multiple performance indicators rather than relying exclusively on accuracy. From a practical perspective, the findings may assist researchers, software developers, digital platform administrators, and content moderation practitioners in selecting classification algorithms that align with their operational priorities. Additionally, the identification of potential lexical biases may support the development of more reliable and context-sensitive misinformation detection systems.

The findings of this study have methodological, technological, and practical implications for automated hoax detection. Methodologically, the differences in model performance indicate that classification accuracy alone is insufficient to determine the suitability of machine learning algorithms for misinformation detection, particularly when the dataset exhibits substantial class imbalance. The higher recall achieved by Naive Bayes suggests its potential usefulness in applications prioritizing the detection of misinformation, whereas the higher precision of Logistic Regression indicates its potential relevance when minimizing false accusations against legitimate content is essential. Meanwhile, the superior accuracy and F1-score of Linear SVM demonstrate its comparative effectiveness in balancing overall classification performance within the evaluated dataset. Technologically, the association between debunking-related vocabulary and valid classifications highlights the possibility that TF-IDF-based models may rely on superficial linguistic indicators rather than deeper semantic understanding. These findings suggest the importance of incorporating contextual language representations and evaluating model generalizability across different writing styles, platforms, and periods. Practically, the study emphasizes that automated hoax detection systems should be designed with consideration of moderation objectives, classification errors, and the potential consequences of incorrectly labeling legitimate information. However, these implications should be interpreted within the limitations of the dataset and experimental design, as the findings do not establish model performance across different social media environments or demonstrate the effectiveness of the algorithms in real-world deployment.

METHOD

This study employed an experimental quantitative approach based on supervised text classification. The data consisted of a public corpus containing 4,617 tweets in Indonesian collected from platform X (formerly Twitter) by a research team at the University of Muhammadiyah Ponorogo in 2020. The tweets were labeled as hoax, valid, or uncertain (?). After removing tweets labeled as uncertain and those with blank text, 3,754 tweets remained for analysis, consisting of 3,041 tweets labeled as hoaxes (80.98%) and 713 labeled as valid (19.02%). This distribution reflected the class imbalance commonly found in real-world misinformation data.

The text preprocessing stages included: (1) case folding, which converted all text to lowercase; (2) removal of links (URLs) using regular expressions; (3) removal of non-alphabetical characters, such as numbers, punctuation, and symbols; and (4) removal of excess spaces. The cleaned text was then represented numerically using Term Frequency–Inverse Document Frequency (TF-IDF) weighting with a combination of unigram and bigram features. The features were limited to the top 3,000 based on a minimum document frequency of two. The data were then divided into 80% training data (3,003 tweets) and 20% test data (751 tweets) using stratified partitioning to maintain the original class proportions.

Three classification algorithms were compared: (1) Multinomial Naive Bayes, a probabilistic algorithm commonly used as a baseline for text classification; (2) Logistic Regression with balanced class weighting (`class_weight='balanced'`) to reduce bias toward the majority class; and (3) Linear Support Vector Machine with balanced class weighting, calibrated using Platt scaling through three-fold cross-validation to produce probability estimates. All models were implemented in Python 3 using the scikit-learn library with a fixed random seed (42) to ensure the reproducibility of the results.

Model performance was evaluated on the test data using five standard measures: accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). Due to the imbalanced class distribution, recall and precision were analyzed separately rather than relying solely on overall accuracy, as accuracy can be misleading when applied to imbalanced data. In addition, the Logistic Regression coefficients were analyzed to identify the 15 terms (unigrams and bigrams) most strongly associated with each class, with the aim of uncovering the linguistic patterns underlying the model's classification decisions.

The study was limited by the fact that its unit of analysis consisted of short tweets derived from data collected in 2020. Therefore, the linguistic patterns identified may not fully represent the styles of hoax dissemination found on different platforms or during different time periods.

RESULTS AND DISCUSSION

The analysis was carried out on 3,754 tweets in Indonesian using three text classification algorithms. Table 1 presents the results of comparing the performance of the three models on the test data.

Table 1. Comparison of the performance of the text classification model

Models	Accuracy	Precision	Recall	F1 Score	AUC
Naive Bayes	0.8282	0.8267	0.9967	0.9038	0.7742
Logistic Regression	0.7870	0.9118	0.8158	0.8611	0.8156
Linear SVM	0.8429	0.8441	0.9885	0.9106	0.7926

As shown in Table 1, Linear SVM achieved the highest accuracy (0.8429) and the highest F1-score (0.9106). However, Naive Bayes achieved the highest recall (0.9967), almost detecting all hoax tweets in the test data, with lower precision consequences (0.8267) because it also marked the number of valid tweets as hoaxes. In contrast, Logistic Regression achieved the highest precision (0.9118) and highest AUC (0.8156), but with a more moderate recall (0.8158). This pattern is consistent with the precision-recall trade-offs that are common in the classification of texts with unbalanced classes (Fahrur et al., 2025).

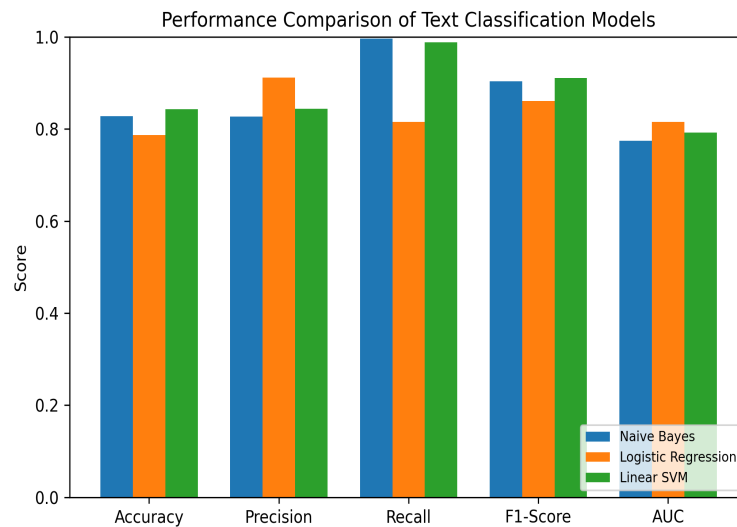


Figure 1. Comparison of accuracy, precision, recall, F1-score, and AUC on three text classification models.

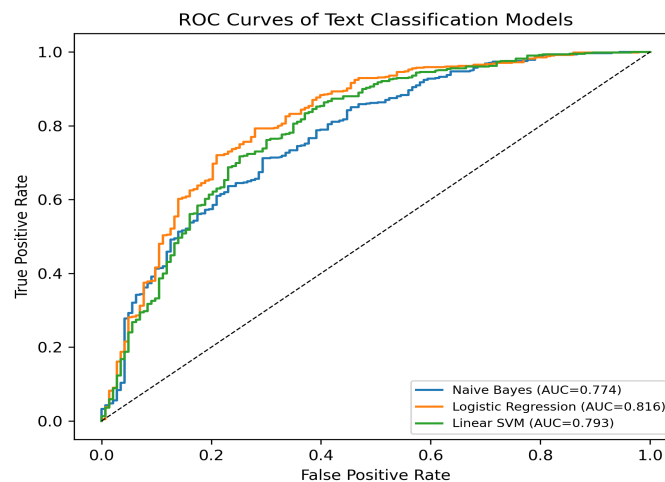


Figure 2. ROC curve of three text classification models on test data.

From an operational perspective, these findings imply that the selection of models for hoax content moderation systems should be tailored to platform priorities. Where the top priority is to minimize hoaxes that escape detection (avoiding false negatives), a model with high recall such as Naive Bayes is more appropriate, although it risks flagging more valid content as hoaxes. Conversely, if the top priority is to avoid mislabeling valid content (avoiding false positives), which can lead to unsubstantiated accusations or even silencing of correct information, high-precision models such as Logistic Regression are more appropriate. These findings expand on the results of Fernandes and Shita (2024) who compared SVM and Random Forest for hoax detection without explicitly discussing the implications of precision-recall trade-offs on content moderation policies.

Further analysis of the terms with the highest Logistic Regression coefficient revealed an interesting and methodologically important finding.

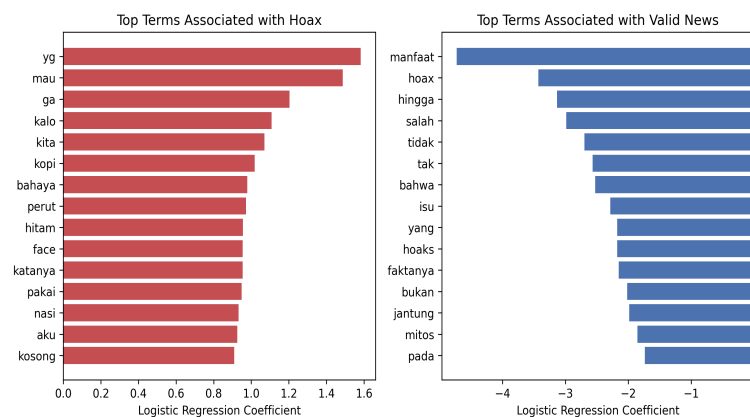


Figure 3. Terms with the strongest association with hoax class (left) and valid class (right) based on the Logistic Regression coefficient.

As shown in Figure 3, the terms that are most associated with the valid class are dominated by vocabulary related to debunking, such as "hoax", "hoax", "false", "fact", "not", and "myth". This happens because tweets labeled valid in this corpus mostly come from accounts or content that explicitly clarify that an issue is a hoax, so the tweet actually mentions the word "hoax" directly, while tweets labeled hoaxes themselves rarely refer to themselves as hoaxes. In contrast, terms associated with the hoax class tend to be everyday informal vocabulary ("yg", "ga", "kalo", "we", "me") and health-lifestyle-topic-specific terms ("coffee", "stomach", "rice", "danger"), which reflect the characteristics of viral content in the style of popular health claims, consistent with the pattern observed by Setiawan et al. (2021) on the detection of news validity on Twitter.

These findings have important methodological implications: models trained on this kind of data run the risk of learning the debunking writing style as a valid class marker, rather than actually learning the factual truth of the content itself. If the model is applied to the original tweet (not the debunking tweet) that addresses a similar topic without explicitly mentioning the word "hoax", the model's performance is likely to decline significantly. This kind of risk is also hinted at by Azizah et al. (2023) and Tobing et al. (2025), who emphasize that simple lexical-based models are susceptible to surface-level bias and recommend context-based transformer models such as IndoBERT to capture deeper semantic meanings, rather than relying solely on the occurrence of specific keywords.

This study has a number of limitations. First, the corpus used is derived from 2020 data and is relatively small (3,754 tweets after cleanup), so generalizations to other platforms or time periods need to be done carefully. Second, the TF-IDF approach is lexical and does not capture the semantic context like transformer models such as BERT or IndoBERT, which according to Azizah et al. (2023) can achieve higher accuracy (87.6%) than a simple lexical feature-based approach. Third, the findings on debunking vocabulary bias suggest that the evaluation of the hoax detection model needs to consider generalizations across writing styles, not just performance metrics in the test data derived from the same distribution as the training data. Further research is recommended using transformer-based models, extending the scope of the data to different platforms and time periods, and testing model generalizations on data that intentionally excludes explicit debunking vocabulary.

CONCLUSION

The study compared three text classification algorithms, namely Naive Bayes, Logistic Regression, and Linear SVM, to detect Indonesian-language hoaxes on social media data. The results showed that no single algorithm excelled across all metrics: Linear SVM achieved the highest accuracy and F1-score, Naive Bayes achieved the highest recall, and Logistic Regression achieved the highest precision and AUC, so model selection needed to be tailored to the operational priorities of the content moderation platform. Further term analysis revealed that the model tends to associate debunking vocabulary with valid classes, a finding that highlights the important limitations of purely lexical-based classification models and the importance of considering model generalizations beyond the specific writing style of the training data. From a socio-technological scientific perspective, this study emphasizes that the development of an effective hoax detection system requires not only optimization of performance metrics, but also an in-depth understanding of the social context and linguistic patterns of the spread of misinformation in Indonesia.

FUNDING

The author(s) declared that financial support was not received for this work and/or its publication.

CONFLICT OF INTEREST

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

If applicable, the Author's Contribution Statement.

GENERATIVE AI STATEMENT

The author(s) declared that generative AI was used in the creation of this manuscript. The author(s) wish to acknowledge the use of Generative AI for refining the language clarity and grammatical structure of this manuscript. This application was conducted in adherence to established ethical guidelines regarding the use and disclosure of generative AI in scholarly research. Each author has provided substantial intellectual contributions to the work, which has been rigorously vetted for accuracy, and the author(s) assume full responsibility for the integrity and final content of the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

REFERENCES

- Agustina, N., Adrian, A., & Hermawati, M. (2022). Implementation of the Naïve Bayes Classifier Algorithm to Detect Fake News on Social Media. *Exacta Factor*, 14(4), 206–213.
- Al Fachrozi, M., & Tania, K. D. (2025). Comparison of Naïve Bayes, SVM, K-NN, Decision Tree, and Random Forest in Sentiment Analysis Based on SeaBank Application Aspects. *Jurnal Teknologi Dan Sistem Informasi*, 12(1), 69–80.

- Ariatmanto, D., & Rifai, A. M. (2024). The Impact of Feature Extraction in Random Forest Classifier for Fake News Detection. *RESTI Journal (Information Systems and Technology Engineering)*, 8(6), 730–736.
- Azizah, S. F. N., Sihwi, S. W., Cahyono, H. D., & Widiarto, W. (2023). Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection. *ArXiv Preprint*. <https://doi.org/10.48550/arXiv.2308.04950>
- Bagus, D., Prasetyo, C., Andono, N., & Supriyanto, C. (2023). Naive Bayes Classifier and Forward Selection Method for the Detection of Indonesian Hoax News. *Journal of Media Informatics Budidarma*, 7(3), 1541–1550.
- Bashaddadh, O., Omar, N., Mohd, M., & Khalid, M. N. A. (2025). Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements. *IEEE Access*, 13, 90433–90466. <https://doi.org/10.1109/ACCESS.2025.3572051>
- Elisabeth, G., Bilah, R. S., Ardini, S. N., Agustina, N., & Rismayadi, D. A. (2023). Classification of Fake News on Fuel Oil (BBM) Price Increase Using Support Vector Machine (SVM) Algorithm. *Narrative: National Journal of Informatics Research, Applications and Techniques*, 5(2), 104–109.
- Fahrur, I., Maulidia, I., Hani, M., Arianto, R., Anantaf, D. R., & Yuli, Y. A. (2025). Comparison of Feature Extraction in Support Vector Machine (SVM) Based Sentiment Analysis System. *Jurnal Ilmiah Kursor*, 13(1), 1–12.
- Fernandes, R. P., & Shita, R. T. (2024). The Application of SVM and Random Forest Methods to Detect Hoax News at PT. Global Arrow. *Ticom Journal: Technology of Information and Communication*, 12(3), 102–107.
- Mridha, M. F., Keya, A. J., Hamid, M. A., Monowar, M. M., & Rahman, M. S. (2021). A comprehensive review on fake news detection with deep learning. *IEEE Access*, 9, 156151–156170. <https://doi.org/10.1109/ACCESS.2021.3129329>
- Nabil, A. A., Wildantama, F. R., Satrianto, D., Bakara, M. G., Budiawan, I., & Mulyat, D. (2026). Hoax Classification Using the TF-IDF + SVM Method. *Journal of Community Service and Educational Research*, 4(3), 17653–17660.
- Najwan, T. A., Karrar, F. H., Muataz, N. A., & Zainab, S. A. (2024). The Application of Random Forest to the Classification of Fake News. *BIO Web of Conferences*, 97, 00049.
- Nurhikam, A. S., Syaputra, R., & Rohman, S. (2023). Detection of Fake News in the 2024 Election Using Random Forest Algorithms. *DoubleClick: Journal of Computer and Information Technology*, 7(1), 41–50.
- Ridho, M. Y., & Yulianti, E. (2024). From text to truth: Leveraging IndoBERT and machine learning models for hoax detection in Indonesian news. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)*, 10(3), 544–555. <https://doi.org/10.26555/jiteki.v10i3.29450>
- Setiawan, E. I., Johaness, S., Hermawan, A. T., & Yamasari, Y. (2021). Detection of News Validity on Twitter Social Media with Naive Bayes Algorithm. *Journal of Intelligent System and Computation*, 3(2), 55–60.
- Tobing, C. J. L., Wijayakusuma, I. G. N. L., & Ida, L. P. (2025). Comparison of IndoBERT and mBERT Performance for the Detection of Political Hoax News in Indonesian. *Journal of Science and Technology*, 14(1), 114–123.