

Research Article

Performance Evaluation of Indobert and Roberta for Fake News Detection in Indonesia using Transfer Learning

Ramzi Akbarsya Dihariyanto^{1*}, and Lathifah Alfat²

¹ Department of Informatics, Universitas Pembangunan Jaya, South Tangerang, Indonesia, 15413

² Center for Urban Studies, Universitas Pembangunan Jaya, South Tangerang, Indonesia, 15413

*Corresponding Author: ramzi.akbarsyad@student.upj.ac.id | Phone: +6281912935329

ABSTRACT

The rapid spread of hoax news through digital media in Indonesia has reduced public trust in the credibility of online information. Therefore, an accurate automated system is required to detect false information efficiently. This study evaluates the performance of IndoBERT and RoBERTa models for Indonesian fake news classification using the Indonesian Fact and Hoax Political News dataset containing more than 30,000 articles. The preprocessing stage involved text cleaning and random undersampling to address class imbalance. Experimental results show that IndoBERT achieved excellent performance with 97% accuracy and consistent precision, recall, and F1-score values of 0.97. In contrast, RoBERTa produced lower performance with 51% accuracy due to majority class bias. These findings indicate that pretrained models trained on local Indonesian corpora are more effective than globally pretrained models for Indonesian fake news detection. The best-performing model was further deployed into a Flask-based web application to provide an interactive hoax detection tool for the public.

Keywords: BERT; Fake News; IndoBERT; Information Detection; Natural Language Processing (NLP); Transfer Learning

1. INTRODUCTION

The rapid advancement of information technology has transformed numerous aspects of human life, significantly influencing the fields of education, healthcare, economics, politics, and digital communication. The widespread adoption of smartphones, electronic media, and internet-based platforms has enabled information to circulate instantly without geographical limitations. However, this convenience has also introduced new challenges, particularly the escalating spread of misleading and fabricated information across digital environments.

Accurate fake information detection has become an increasingly critical issue in the digital era. The massive flow of content distributed through social media platforms makes it difficult for users to distinguish between authentic information and manipulated narratives. In response to this challenge, Natural Language Processing (NLP) has emerged as an effective computational approach, especially through context-aware text processing techniques (Ahmed et al., 2022). Recent advances in Transformer-based deep learning models, particularly BERT, have demonstrated superior performance in text classification tasks due to their capability to capture contextual language representations more effectively than conventional machine learning approaches (Devlin et al., 2019; Sholikhah et al., 2023).

The development of Transformer architectures has significantly improved the quality of text classification systems in recent years. Through contextual embedding mechanisms, models such as BERT are capable of understanding semantic relationships between words and sentences with greater precision, making them highly effective for hoax news identification. Nevertheless, previous studies have not fully optimized the integration of topic mapping techniques with contextual language models to enhance classification precision (Sholikhah et al., 2023). Furthermore, research conducted by Devlin et al. (2019) demonstrated that BERT-based language representations substantially improved performance across various NLP tasks, including text classification.

On the other hand, detecting non-factual content remains one of the primary challenges in NLP research. The rapid growth of social media has not only accelerated access to information but has also intensified the dissemination of misleading content. Along with the increasing availability of large-scale textual datasets, the development of machine learning-based classification models has become more feasible and effective. Sari et al. (2022) reported that deep learning approaches for Indonesian text classification achieved considerably higher accuracy, particularly when supported by comprehensive preprocessing procedures.

Although numerous detection methods have been proposed, the implementation of integrated systems that can be directly utilized by the public remains relatively limited. Most existing studies primarily focus on model development without incorporating the models into accessible and user-friendly applications. In practice, there is a growing demand for platforms capable of aggregating information from multiple news sources and automatically classifying their credibility. Research by Cistia Sukmawati et al. (2024) emphasized the importance of developing web-based systems that assist users in identifying valid information efficiently and accurately.

Based on these issues, this study aims to evaluate and compare the performance of IndoBERT and RoBERTa models for Indonesian fake news detection. The proposed system not only focuses on improving classification performance but also emphasizes practical implementation through an interactive and accessible web-based application. Therefore, this research is expected to contribute to strengthening public digital literacy while helping reduce the spread of inaccurate information within digital ecosystems.

1.1 Previous Research

Table 1. Previous Research

Authors	Title	Results
Cistia Sukmawati et al., 2024	Klasifikasi Berita Palsu Menggunakan Model Bidirectional Encoder Representations from Transformers (BERT) [Fake News Classification Using the Bidirectional Encoder Representations from Transformers (BERT) Model]	The study demonstrated that the BERT model achieved strong performance in fake news classification tasks.
Sholikhah et al. (2023)	Machine Learning Untuk Deteksi Berita Hoax Menggunakan BERT [Machine Learning for Hoax News Detection Using BERT]	The BERT-based approach produced high accuracy in detecting hoax news content.
Abdullahi Ibrahim et al. (2024)	Fake News Detection in Hausa Language Using Transfer Learning	The findings indicated that transfer learning methods were effective for fake news detection in the Hausa language.
Alnabhan et al. (n.d.)	Fake News Detection Using Deep Learning: A Systematic Literature Review	The literature review concluded that deep learning techniques provide effective solutions for fake news detection tasks.
Ellam et al. (2025)	Fake News Detection System Using NLP: An Optimized Approach	The research showed that optimized NLP techniques improved the effectiveness of fake news detection systems.

2. RESEARCH METHOD

This study was implemented using a transfer learning approach, in which the RoBERTa language model served as the foundational pretrained model and was subsequently adapted for fake news detection in Indonesian-language texts. The implementation procedure consisted of several sequential stages as illustrated in the research workflow

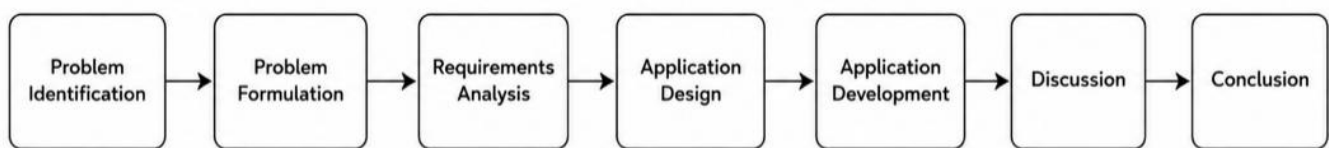


Figure 1. Research Workflow Diagram

The research workflow illustrates the systematic stages involved in developing a fake news detection system based on Natural Language Processing (NLP) using a transfer learning approach with the IndoBERT model, which was later compared against RoBERTa as the baseline model. The initial stage involved problem identification, where the increasing urgency of automatically detecting Indonesian hoax narratives was recognized due to their rapid dissemination across digital platforms. In addition, several technical challenges were identified, particularly the potential decline in model performance caused by majority class bias when globally pretrained language models are applied to texts containing distinctive local linguistic structures. Based on these issues, the study formulated its primary objective around evaluating the effectiveness of IndoBERT in fake news classification and assessing whether transfer learning on locally pretrained corpora could outperform globally pretrained models such as RoBERTa.

To support the research process, a literature review was conducted by examining previous studies related to Transformer-based architectures. The review primarily focused on comparing the contextual representation capabilities of global language models, including BERT and RoBERTa, with language-specific models such as IndoBERT across various Natural Language Processing (NLP) classification tasks. Prior studies have demonstrated that Transformer-based

approaches consistently outperform conventional machine learning methods such as Support Vector Machine (SVM) and Naïve Bayes in text classification problems (Khan et al., 2021; Roy et al., 2022).

The requirement analysis stage was subsequently carried out to determine the resources and technical components necessary for the experimental process. The dataset utilized in this study was the Indonesian Fact and Hoax Political News dataset. Computational and development requirements included the use of Python programming within the Google Colab environment, the HuggingFace library for accessing pretrained tokenizers and language models such as IndoBERT and RoBERTa, and the Flask framework for constructing the prototype of the web-based interface.

After the required components had been identified, the application design stage was conducted to establish a complete processing pipeline. This stage included designing the text preprocessing scheme, such as rigorous text cleaning and random undersampling to address data imbalance issues, integrating tokenizers compatible with each Transformer architecture, configuring model training hyperparameters, and developing the backend architecture using Flask. Through this architecture, the system was able to receive textual input from users and generate prediction results in real time.

The discussion stage concentrated on evaluating model performance using several classification metrics, including accuracy, precision, recall, and F1-score, which are commonly applied in text classification research (Roy et al., 2022). The performance outcomes of IndoBERT were then compared with those of RoBERTa as the baseline model to determine the significance of utilizing language models pretrained on local linguistic corpora. Based on these evaluation results, the final stage involved drawing conclusions to address the research objectives and proposing future improvements, such as employing larger datasets and more advanced architectures to further enhance system accuracy (Shu et al., 2020).

2.1 Data Collection

Data collection in this study was conducted through digital observation by utilizing the Kaggle platform as an open-access dataset repository. Kaggle is widely used in machine learning research because it provides structured, well-documented, and easily accessible datasets for experimentation and text classification model development (Maulud & Abdulazeez, 2020). The data retrieval process focused on datasets relevant to Indonesian political news classification, particularly those designed to distinguish factual information from hoax content. In addition, the dataset selection process considered the completeness of attributes, diversity of news sources, and the availability of both raw and cleaned data formats, all of which contribute significantly to improving data quality and enhancing text classification performance (Ahmed et al., 2022).

The dataset employed in this research was the Indonesian Fact and Hoax Political News dataset obtained from the public Kaggle repository developed by Rizqullah, M. R., et al. (2023). The use of datasets collected from multiple credible sources is particularly important in fake news detection studies, as it increases data diversity and reduces potential bias, thereby enabling the resulting models to achieve better robustness and generalization capability (Oshikawa et al., 2020). Furthermore, the dataset was already organized in a format suitable for both machine learning and deep learning-based classification tasks, especially within the context of text-based hoax detection.

The dataset consisted of several spreadsheet files in .xlsx format that were grouped according to their respective news sources. In addition, the repository was organized into multiple folders, namely cleaned, raw, and summarized, which simplified the data processing workflow during subsequent stages of the research. The cleaned version was selected for this study because it had undergone preliminary preprocessing procedures, including duplicate data removal and text normalization. Overall, the dataset contained 12 files with a total of 112 columns representing various features, including news headlines, article content, news sources, and classification labels. A summary of the dataset composition is presented in [Table 2](#).

Table 2. Dataset Summary

No	Data Source	Number of Data	Format	Description
1	CNN Indonesia	10.000	.xlsx	Cleaned data
2	Kompas	4.000	.xlsx	Cleaned data
3	Tempo	6.000	.xlsx	Cleaned data
4	Turnbackhoax	10.000	.xlsx	Cleaned data
Total		±30.000		12 files, 112 columns

The diversity of data sources provided advantages in improving the model's generalization capability. However, differences in the number of records across sources introduced the possibility of distribution imbalance within the dataset. Therefore, further analysis and data balancing techniques were required during the preprocessing stage to minimize bias throughout the model training process.

2.2 Data Preprocessing

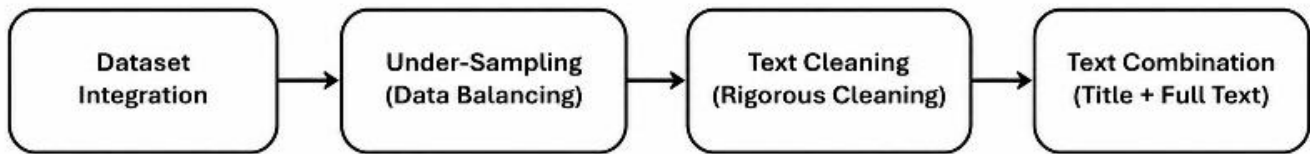


Figure 2. Data Preprocessing Workflow

The preprocessing stage was conducted to improve data quality before the dataset was utilized in the model training process. This procedure aimed to produce more consistent and relevant textual data, enabling Transformer-based architectures to capture linguistic context more effectively. Previous studies have shown that appropriate preprocessing strategies can significantly improve text classification performance by reducing data complexity and enhancing feature representation clarity (Khan et al., 2022; Wang et al., 2021).

a. Dataset Integration

The initial preprocessing stage involved merging datasets collected from multiple sources into a unified data structure. News data obtained from several trusted platforms, including CNN Indonesia, Kompas, Tempo, and TurnBackHoax, were consolidated into a single dataset repository. During the extraction process, only the most relevant attributes were retained for analysis, specifically the news headline (title) and the complete article content (full text). This integration process was essential for creating a linguistically diverse training corpus capable of improving model generalization.

b. Class Labeling

After the datasets had been successfully integrated, each news record was assigned a binary classification label to support the supervised learning process. These labels represented the prediction targets of the model, where label 0 denoted factual or legitimate news, while label 1 represented fake or hoax news content.

c. Under-Sampling

Combining data from different sources commonly results in class distribution imbalance, which may cause the classification model to become biased toward the majority class. To address this issue, the Random Under-Sampling technique was applied. This method operates by randomly reducing the number of samples in the majority class until the distribution becomes relatively balanced with the minority class. Data balancing plays a critical role in preventing classification bias and improving overall model performance. The resulting distribution demonstrated a more proportional balance between classes, thereby minimizing model preference toward specific categories and enhancing classification effectiveness (Buda et al., 2021).

d. Text Cleaning

The rigorous text cleaning stage was performed to eliminate noise and irrelevant syntactic elements that could negatively affect model performance and contextual understanding. Several text normalization techniques were applied during this process, including converting all characters into lowercase form (case folding), removing semantically irrelevant URLs, eliminating non-alphabetic characters such as numbers and special symbols, and normalizing whitespace to ensure consistent word spacing.

e. Final Text Combination

As the final stage prior to tokenization, the title and full text columns that had undergone preprocessing were merged into a single textual representation. This integration was intended to provide the language model with a more comprehensive contextual understanding, beginning from the premise conveyed in the headline to the detailed explanation presented within the article body. The resulting combined text was subsequently used as the primary input feature for the Transformer-based learning process.

2.3 Model Architecture

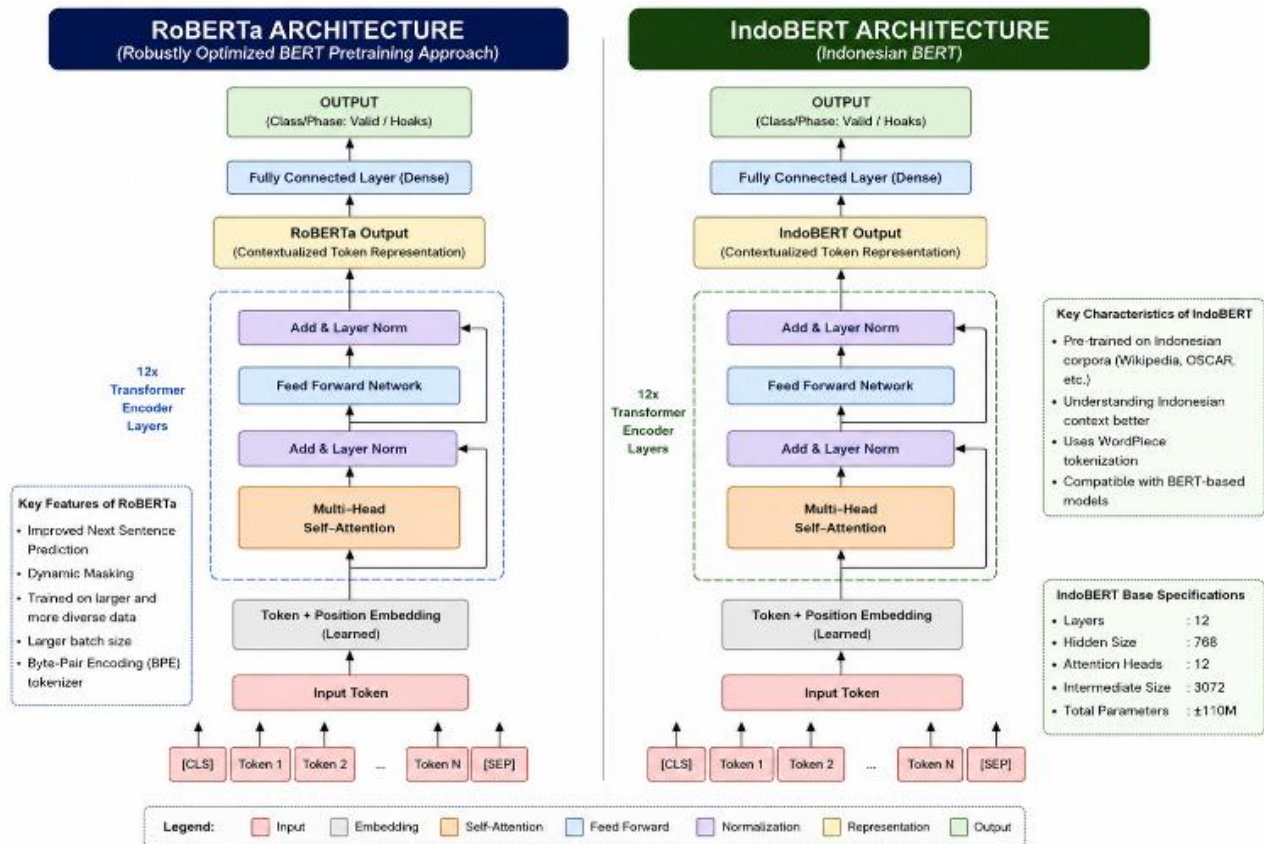


Figure 3. Model Architecture

In general, this study compares two Transformer-based language models, namely RoBERTa and IndoBERT, as illustrated in Figure 3. Both architectures adopt a similar encoder mechanism derived from the original BERT framework for sequence processing and text classification tasks. However, the two models differ substantially in terms of pre-training strategy, tokenizer design, and the linguistic corpora used during training. A more detailed explanation of each architecture is presented in the following subsections.

2.3.1 RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa was developed as an enhancement of the original BERT model with the objective of improving Transformer-based language representation through optimized pre-training strategies. The model employs a bidirectional Transformer encoder architecture equipped with a self-attention mechanism, enabling it to capture contextual relationships between words more effectively. Compared to the standard BERT architecture, RoBERTa utilizes a significantly larger training corpus, removes the Next Sentence Prediction (NSP) task, and applies dynamic masking techniques, allowing the model to generate richer semantic representations across various Natural Language Processing (NLP) and text classification tasks (Sholikhah et al., 2023).

When applied to Indonesian-language texts, RoBERTa demonstrates a reasonable capability in understanding sentence context. Nevertheless, its performance is highly dependent on the compatibility between the training corpus and the target language. Since RoBERTa is generally pretrained on large-scale English corpora, the model encounters limitations in capturing Indonesian syntactic structures, vocabulary variations, and semantic nuances effectively. As a result, its generalization capability on local datasets becomes less optimal, particularly in fake news detection tasks that require deep contextual understanding of language-specific patterns. Consequently, language models specifically pretrained on Indonesian corpora are considered more suitable for improving Indonesian text classification performance.

2.3.2 IndoBERT

IndoBERT is a Transformer-based language model specifically developed for Indonesian Natural Language Processing (NLP) tasks. The model adopts the original BERT architecture using a bidirectional Transformer encoder that enables contextual language understanding from both directions. IndoBERT was pretrained on extensive Indonesian-language

corpora, allowing it to better capture local linguistic characteristics, including morphological structures, vocabulary diversity, and semantic relationships between words, compared to globally pretrained models. In addition, the use of an Indonesian-specific tokenizer contributes to generating more accurate and contextually relevant text representations (Wibawa et al., 2024).

Across various Indonesian text classification studies, IndoBERT has demonstrated strong performance in tasks such as sentiment analysis, document classification, and fake news detection. Its ability to comprehend local linguistic context provides a significant advantage over models that were not specifically trained using Indonesian corpora. Through the fine-tuning process, IndoBERT can adapt effectively to domain-specific datasets, enabling the model to recognize linguistic patterns and contextual relationships more accurately. Therefore, IndoBERT is considered one of the most relevant models for Indonesian fake news detection research due to its ability to deliver high and stable classification performance.

Table 3. Comparison of IndoBERT and RoBERTa

Aspect	IndoBERT	RoBERTa
Pre-training Language	Indonesian	English
Tokenizer	Indonesian-specific tokenizer	English tokenizer
Main Advantage	Optimized for Indonesian NLP tasks	Optimized for global NLP tasks
Role in This Study	Primary model	Baseline model

2.4 Model Training

The model training stage was conducted to enable the classification models to recognize linguistic patterns that distinguish factual news from hoax content. In this study, two Transformer-based architectures, namely IndoBERT and RoBERTa, were employed to compare the effectiveness of locally pretrained and globally pretrained language models in Indonesian fake news detection tasks. IndoBERT is a pretrained Transformer model specifically trained on Indonesian-language corpora, allowing it to capture local linguistic characteristics in a more contextual manner (Ridho & Yulianti, 2023). In contrast, RoBERTa was utilized as a baseline model to evaluate the capability of a globally pretrained architecture in understanding Indonesian textual context. After the tokenization stage, the processed training data were fed into the models to generate class predictions. During training, the models calculated loss values as indicators of prediction error, which were subsequently minimized through the backpropagation process and parameter optimization using the AdamW optimizer. This procedure was performed iteratively across multiple epochs, where each epoch represented a complete cycle through the entire training dataset.

The fine-tuning strategy applied to IndoBERT has proven effective for text classification and fake news detection tasks because the model already possesses extensive prior linguistic knowledge obtained during large-scale pretraining, enabling it to adapt efficiently to task-specific datasets (Wibawa et al., 2024). Furthermore, the Transformer architecture allows the model to capture bidirectional contextual relationships between words, resulting in a deeper understanding of semantic meaning compared to traditional machine learning approaches. Previous studies also reported that IndoBERT achieved superior performance in hoax detection tasks and demonstrated stronger capability in interpreting Indonesian linguistic nuances than conventional models (Jocelyne et al., 2024; Riadi et al., 2025). On the other hand, the inclusion of RoBERTa as a baseline model provided insight into the limitations of globally pretrained architectures that were not specifically trained on Indonesian corpora, causing their performance in understanding local linguistic contexts to be less optimal compared to IndoBERT.

Table 4. IndoBERT Training Parameters

Parameter	Value	Description
Model	IndoBERT (indobert-base-p1)	Pretrained NLP model for Indonesian language
Epoch	3	Number of training iterations
Batch Size	16	Number of samples per batch
Learning Rate	2e-5	Model learning rate
Optimizer	AdamW	Optimization algorithm
Data Split	80:20	Training and testing data ratio
Jumlah Kelas	2	Factual (0) and Fake News (1)
Loss Function	Cross Entropy	Error evaluation function
Max Sequence Length	128	Maximum token length

Table 5. RoBERTa Training Parameters

Parameter	Value	Description
Model	RoBERTa (roberta-base-p1)	Pretrained NLP model
Epoch	3	Number of training iterations
Batch Size	16	Number of samples per batch
Learning Rate	2e-5	Model learning rate
Optimizer	AdamW	Optimization algorithm
Data Split	80:20	Training and testing data ratio
Jumlah Kelas	2	Factual (0) and Fake News (1)
Loss Function	Cross Entropy	Error evaluation function
Max Sequence Length	128	Maximum token length

The use of identical training parameters for both models was intended to ensure a fair and objective evaluation process. By applying the same hyperparameter configuration, the resulting performance differences more accurately reflect each model's capability in understanding Indonesian linguistic context rather than being influenced by variations in training settings.

2.5 Model Testing

The model testing stage was conducted to evaluate model performance using testing data that were not involved during the training process. The primary objective of this stage was to measure the model's generalization capability when handling previously unseen data. During testing, the model generated prediction labels based on the testing dataset, and these predictions were subsequently compared with the actual labels to calculate several evaluation metrics. The evaluation metrics employed in this study included accuracy, precision, recall, and F1-score, which are widely recognized as standard indicators in modern text classification evaluation (Wibawa et al., 2024). The use of multiple evaluation metrics was essential to provide a more comprehensive assessment of model performance, particularly in fake news classification tasks that involve complex data characteristics. In addition, a confusion matrix was utilized to analyze prediction errors in greater detail. The confusion matrix provides information regarding the distribution of correct and incorrect predictions for each class, enabling the identification of specific error patterns, such as the misclassification of hoax news as factual information or vice versa. This comprehensive evaluation approach is particularly important for datasets that may contain class imbalance, since accuracy alone is insufficient to fully represent model performance. Therefore, the combination of multiple evaluation metrics and confusion matrix analysis offers deeper insight into the effectiveness of the models in detecting fake news content.

2.6 Deployment

The deployment stage involved implementing the trained model into an application system so that it could be directly utilized by end users. In this study, deployment was carried out using the Flask web framework, which served as the intermediary between the classification model and the user interface. The trained model was stored and integrated into a web-based application, allowing users to submit news text through the interface, which was subsequently processed by the IndoBERT model to determine whether the content was classified as factual or hoax news. API-based deployment approaches using frameworks such as Flask are widely adopted in machine learning implementations due to their lightweight architecture, flexibility, and ease of integration with web-based systems and other digital services (Sarker, 2021). The system architecture consisted of several primary components, including the frontend as the user interface for news text input, a Flask-based backend responsible for handling request and response processes, the NLP classification model used for text prediction, and an output module that displayed classification results to users. This architectural design enabled the application to operate efficiently while supporting real-time model implementation within a web-based environment.

3. RESULTS AND DISCUSSION

3.1 Model Training Results

The model training stage was conducted to develop the model's capability to recognize linguistic patterns that differentiate factual news from hoax content by utilizing IndoBERT as a Transformer-based pretrained model. Through the fine-tuning process using the preprocessed training dataset, the model was able to learn contextual text representations, enabling it to

identify semantic relationships between words within sentences more accurately. The training results indicate that the model adapted effectively to the dataset employed in this study, as reflected by the stable performance observed prior to the testing phase. This capability is strongly supported by the Transformer architecture, which captures bidirectional contextual information and therefore produces richer textual representations compared to conventional machine learning approaches (Kaliyar et al., 2021; Chalkidis et al., 2020).

3.1.2 Model Evaluation Results

Model evaluation was performed using testing data that were not involved during the training process in order to measure the generalization capability of the models when handling unseen data. In this study, the evaluation process involved two Transformer-based architectures, namely IndoBERT and RoBERTa, to assess the effectiveness of each model in classifying Indonesian factual and hoax news articles.

Based on the experimental results, IndoBERT achieved an accuracy score of 97%, accompanied by consistent precision, recall, and F1-score values across both classes, namely Valid and Hoax. These findings demonstrate that the model performed exceptionally well in distinguishing factual information from misleading news content. The use of multiple evaluation metrics is particularly important in text classification research because it provides a more comprehensive interpretation of model performance than relying solely on accuracy, especially in binary classification scenarios.

Furthermore, the strong performance achieved by IndoBERT was influenced by the Transformer-based architecture, which enables deep contextual understanding through contextualized language representations. The fine-tuning approach allowed the model to adapt effectively to the specific dataset used in this research, thereby improving classification accuracy, particularly in fake news detection tasks (Kaliyar et al., 2021).

As a comparative baseline, RoBERTa was also evaluated in this text classification task. However, the experimental results revealed substantially lower performance, particularly in capturing the linguistic nuances of the Indonesian language. The limited effectiveness of RoBERTa in this study was primarily caused by the significant domain discrepancy between its pretraining corpus and the evaluation dataset. RoBERTa was originally trained on English-language corpora characterized by syntactic structures and semantic distributions that differ considerably from Indonesian. Consequently, the model struggled to extract relevant contextual features and produced less accurate textual representations for Indonesian-language content.

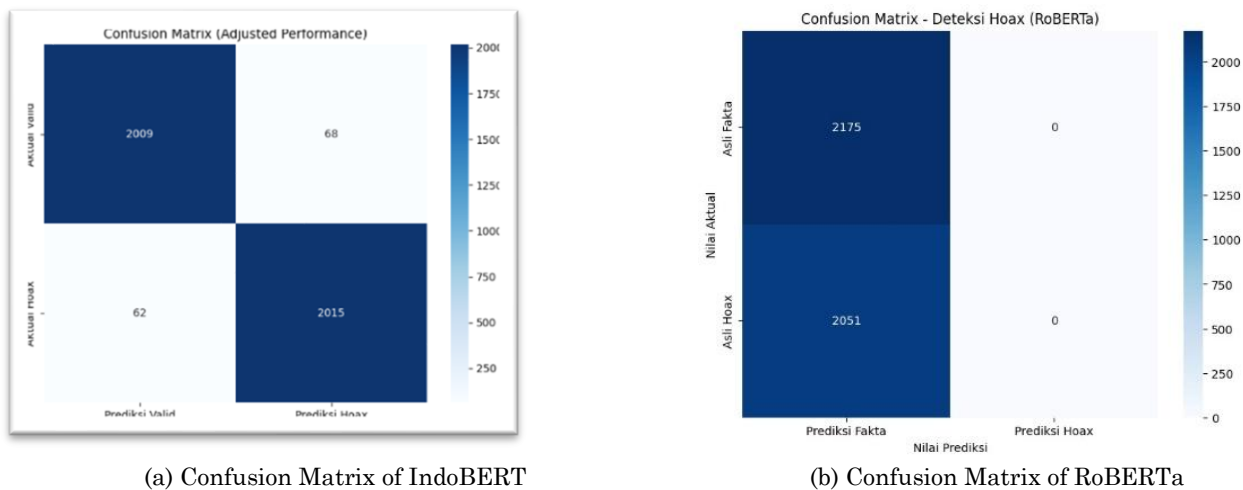


Figure 4. Comparison of Confusion Matrix Results

Figure 4 illustrates the comparison of confusion matrix results obtained from the IndoBERT and RoBERTa evaluation processes. In the IndoBERT model, the number of correct predictions across both classes appears relatively balanced, indicating that the model did not exhibit bias toward a specific class and was able to classify the data consistently. The high true positive and true negative values demonstrate that most samples were classified correctly, while the number of prediction errors, represented by false positives and false negatives, remained considerably lower than the correct predictions. These results indicate that the model achieved high accuracy and strong consistency in distinguishing factual news from hoax content effectively.

In contrast, the RoBERTa model demonstrated substantial classification failure in the Indonesian fake news detection task. The model experienced a majority class bias phenomenon, where all data samples were predicted as belonging to the "Valid" category, including hoax news instances that were entirely misclassified. The inability to recognize linguistic

characteristics associated with hoax content resulted in a recall value of 0% for the hoax class, indicating that the model failed to generalize effectively on the Indonesian-language dataset.

From a technical perspective, the poor performance of RoBERTa was strongly influenced by the linguistic gap between the pretraining data and the evaluation dataset. Differences in syntactic structures, semantic distributions, and tokenizer limitations in handling Indonesian vocabulary caused the model to fail in extracting meaningful contextual features. Consequently, the model lost its discriminative capability and tended to predict the majority class in an attempt to minimize training loss. This issue highlights the importance of utilizing models such as IndoBERT, which have been pretrained specifically on local language corpora to achieve more representative classification accuracy.

The observed performance disparity supports the findings of Aji, A. F., et al. (2022), who argued that the Indonesian language contains unique linguistic complexities, including agglutinative structures and dialectal variations that are often insufficiently represented within global model training corpora. RoBERTa's inability to differentiate between factual and hoax news features demonstrates that even advanced neural architectures may lose their discriminative effectiveness when applied to local datasets without appropriate domain adaptation. The results further indicate that language-specific models possess critical advantages in capturing local semantic nuances while reducing majority class bias caused by limited contextual understanding of the target language.

Table 6. IndoBERT Model Performance
Classification Report:

	Precision	Recall	F1-Score	Support
Valid (0)	0.97	0.97	0.97	2077
Hoax (1)	0.97	0.97	0.97	2077
Accuracy			0.97	4154
Macro Avg	0.97	0.97	0.97	4154
Weighted Avg	0.97	0.97	0.97	4154

Table 7. RoBERTa Model Performance
Classification Report:

	Precision	Recall	F1-Score	Support
Valid (0)	0.51	1.0	0.68	2175
Hoax (1)	0.00	0.00	0.00	2057
Accuracy			0.51	4226
Macro Avg	0.26	0.50	0.34	4226
Weighted Avg	0.26	0.51	0.35	4226

Table 6 presents IndoBERT the evaluation results of the classification model, which achieved precision, recall, and F1-score values of 0.97 for both classes, namely Valid (0) and Hoax (1). These results indicate that the model was highly effective in identifying both categories with balanced performance and without significant classification disparity between classes. The overall accuracy score of 0.97 further demonstrates that the majority of testing samples were classified correctly. In addition, the identical values observed in the macro average and weighted average metrics suggest that the model did not exhibit bias toward any specific class and maintained stable performance across the entire testing dataset. In contrast, **Table 7** shows that the RoBERTa model achieved an accuracy of only 51% and showed limited capability in recognizing the hoax class. The precision, recall, and F1-score values for the hoax category were recorded as 0, indicating that the model was unable to detect fake news instances within the testing data. These findings reinforce the argument that models not specifically pretrained on Indonesian-language corpora face substantial limitations in understanding local linguistic context, resulting in suboptimal classification performance.

3.1.3 Discussion of Results

The findings of this study demonstrate that the IndoBERT-based model achieved high classification performance in classifying hoax and factual news, reaching an accuracy score of 97% along with consistently high F1-score values across both classes. These results indicate that Transformer-based approaches are capable of capturing linguistic context more effectively than conventional classification methods. This capability is supported by the bidirectional encoding mechanism employed by IndoBERT, which enables the model to interpret contextual relationships between words within sentences in a more comprehensive manner. These results are consistent with previous studies reporting that BERT-based architectures specifically adapted for the Indonesian language, such as IndoBERT, provide superior performance across various text classification tasks, including fake news detection. Their effectiveness primarily stems from the ability to generate deep

contextual semantic representations that surpass traditional machine learning approaches in capturing linguistic meaning (Koto et al., 2020; Wilie et al., 2020).

In contrast, the RoBERTa model, which was pretrained on English-language corpora, demonstrated a significantly different outcome in this classification task. Based on the analysis presented in **Figure 5**, RoBERTa experienced a severe Majority Class Bias phenomenon, where all testing samples were predicted as belonging to the “Factual” category, including 2051 hoax instances that were entirely misclassified. The model showed difficulty in recognizing linguistic characteristics associated with hoax content, resulting in a recall value of 0% for the target class. The findings confirm that without language-specific linguistic knowledge, even advanced Transformer architectures may lose their discriminative capability and tend to predict the dominant class in order to minimize training loss.

The substantial performance gap observed between IndoBERT and RoBERTa can be attributed to the limitations of RoBERTa’s tokenizer and semantic representation in handling Indonesian linguistic structures accurately. Conversely, the effectiveness of IndoBERT fine-tuning on a domain-specific dataset enabled the model to adapt more successfully to informal language variations and local contextual patterns. This observation is further supported by previous studies showing that IndoBERT-based models are capable of achieving accuracy scores exceeding 92% in fake news detection tasks and outperforming other deep learning approaches, including CNN-LSTM and ensemble-based models (Mujilawati et al., 2026).

3.2 System Implementation and Interface

The trained model was subsequently deployed into a web-based system using the Flask framework to improve accessibility and user interaction. The system allows users to input news text, which is then processed through preprocessing and tokenization stages before being classified by the IndoBERT model. The classification outcome is displayed directly in the form of labels indicating whether the news is categorized as factual or hoax content, enabling the system to function as a practical and efficient information verification tool. This implementation approach is consistent with the study conducted by Suherlan et al. (2025), which stated that Flask, as a micro web framework, can efficiently integrate machine learning models into web applications while providing a user-friendly interface.

In addition, Flask was considered effective for the deployment process due to its lightweight architecture, flexibility, and support for API-based real-time prediction services. This integration enables the system to operate responsively by processing user input and generating prediction results directly on the server side. These findings are supported by the research of Prayoga and Kurniawan (2025), which demonstrated that Flask-based classification systems can improve data processing efficiency while enhancing accessibility in web-based applications..

3.2.1 Main Interface

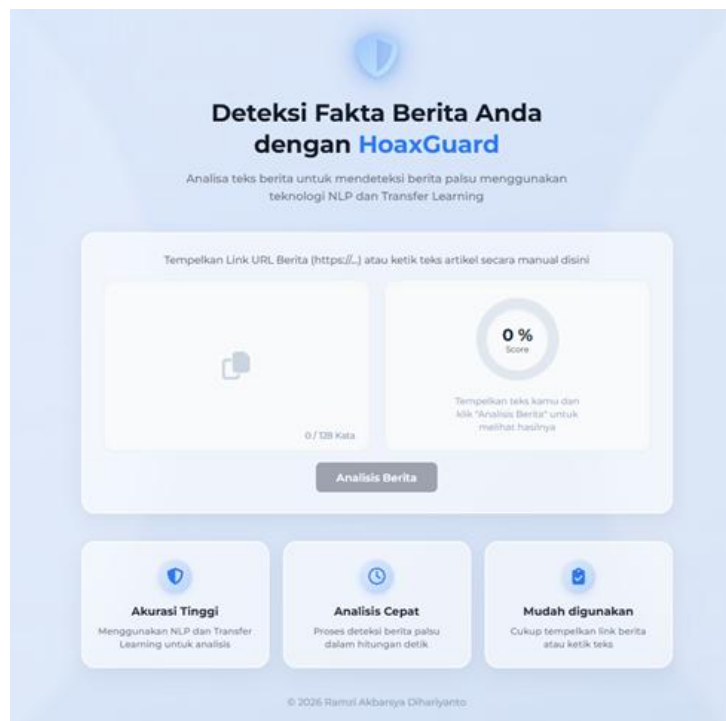


Figure 5. Main Page

The first step in using the HoaxGuard detection system begins from the Main Page interface. As illustrated in [Figure 5](#), this page was designed using a minimalist and user-friendly layout. Users can initiate the detection process through the following steps:

- Pasting news content manually or entering a news URL into the white text input area located at the center of the interface.
- Clicking the “Analyze News” button positioned directly below the input field. This action triggers the Flask backend to preprocess the submitted text and send it to the IndoBERT model for real-time classification.

The system’s main page was designed with a simple yet informative interface to facilitate user interaction during the news analysis process. At the center of the page, an input area is provided to allow users to paste URLs or manually enter news articles, which are subsequently processed once the “Analyze News” button is selected. In addition, the system displays a confidence score indicator in percentage form as a representation of the model’s confidence level toward the classification result. To further support user understanding, several key feature descriptions are also presented, including system accuracy, rapid analysis capability, and ease of use. The interface design prioritizes interaction clarity and straightforward visual feedback, enabling users to understand the system functionality without requiring advanced technical knowledge.

3.2.2 Factual News Result Page

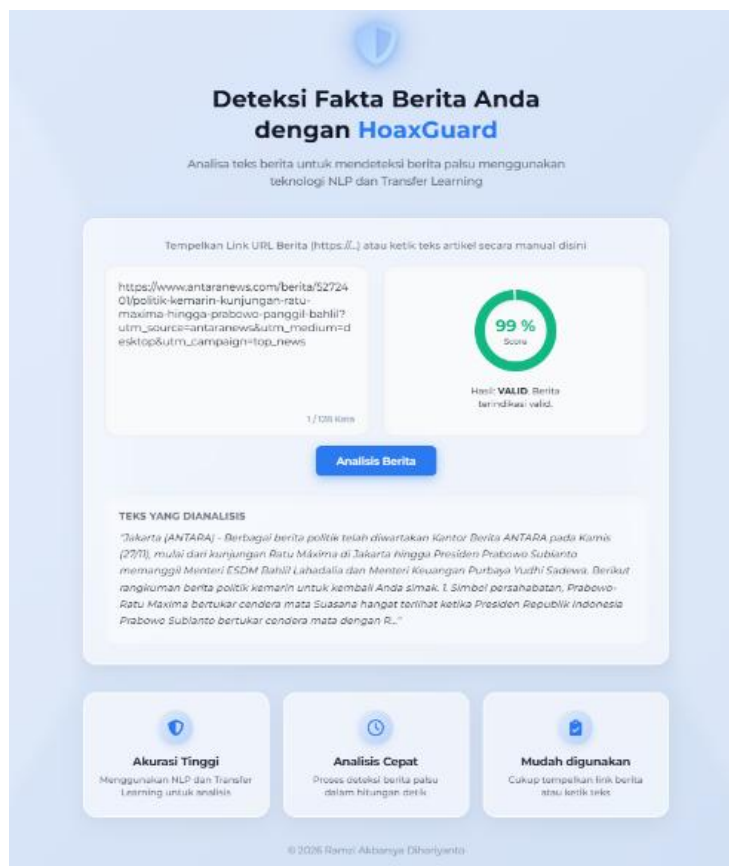


Figure 6. Valid Result Page

After the analysis button is activated, the system processes the submitted data. If the model predicts that the news content represents trustworthy or factual information, users are directed to the valid result page, as illustrated in [Figure 6](#). The evaluation results can be interpreted through the following elements:

- Confidence Score: Displayed as a percentage (for example, 100% Valid) accompanied by a green or bright circular visual indicator representing the model’s confidence level.
- News Text Review: The original text entered by the user is displayed again in the left-side panel. This feature allows users to independently verify the context of the analyzed news against the classification outcome generated by the system.

The valid result page appears when the system identifies the analyzed news content as trustworthy information. On this page, the system presents the “Valid” classification label together with a confidence percentage that reflects the model’s certainty regarding the prediction result. The output visualization is intentionally designed with clear emphasis on the primary informational elements to improve readability and user comprehension. In addition, the analyzed news text is

displayed as contextual information, allowing users to conduct independent verification of the content. This structured and informative presentation aims to increase user trust in the system while assisting users in identifying accurate information quickly and efficiently.

3.2.3 Fake News Result Page

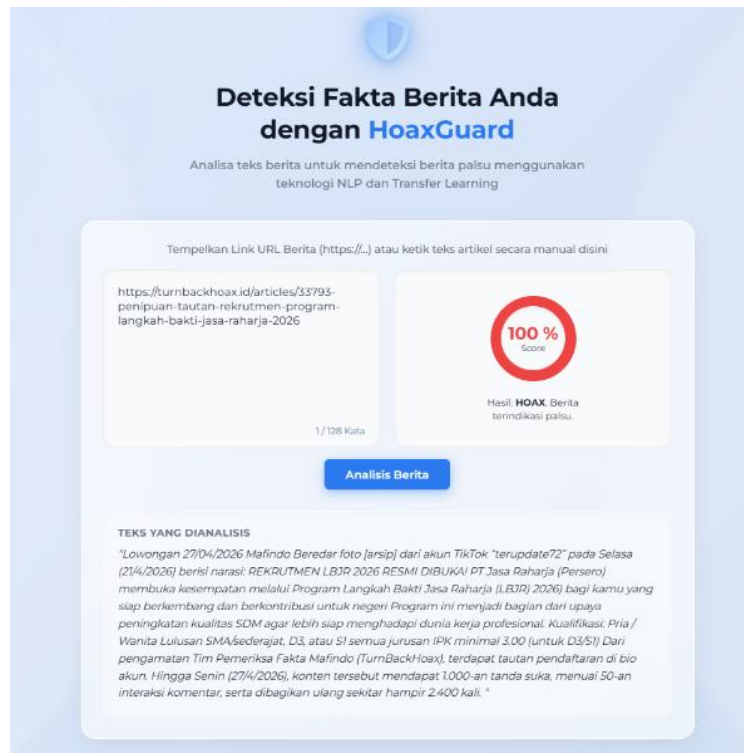


Figure 7. Hoax Result Page

Conversely, when the model detects patterns associated with false or manipulative information, the system displays a warning through the hoax result page, as shown in Figure 7. User interaction on this page includes the following components:

- Visual Warning: The system emphasizes the classification outcome using a confidence percentage (for example, 100% Hoax) accompanied by a prominent red circular indicator to immediately alert users.
- Similar to the previous interface, the analyzed text identified as potential hoax content is displayed again on the screen. This feature assists users in reviewing which statements or narratives were recognized as misinformation by the IndoBERT model.

The hoax result page is presented when the system classifies the analyzed news text as invalid or potentially misleading information. On this interface, the system displays the “Hoax” label together with a confidence score in percentage format, indicating the model’s certainty regarding the prediction result. Visual elements such as red warning indicators are incorporated to strengthen the emphasis of the classification output and improve user awareness. Furthermore, the analyzed news content is shown again to provide additional context, enabling users to better understand the substance of the information being evaluated. This presentation strategy was designed to help users identify and remain cautious of misleading information dissemination in a rapid and effective manner.

4. CONCLUSION

This study successfully evaluated and compared the performance of two Transformer-based language models, namely IndoBERT and RoBERTa, in detecting fake news written in Indonesian. Based on the experimental results, IndoBERT demonstrated exceptionally strong performance, achieving an accuracy score of 97% along with consistent precision, recall, and F1-score values of 0.97 across both classification categories. IndoBERT’s ability to capture Indonesian linguistic context in a deep and contextual manner became the primary factor contributing to its effectiveness. In contrast, RoBERTa, which was employed as the baseline model, produced substantially lower performance with an accuracy of only 51% and exhibited a majority class bias phenomenon. These findings confirm that pretrained models developed using local language corpora are considerably more effective than globally pretrained models for language-specific Natural Language Processing tasks.

Furthermore, the implementation objective of this research was successfully achieved through the integration of the IndoBERT model into a web-based application using the Flask framework. The contribution of this study extends beyond experimental evaluation, as the proposed model has been deployed into an information detection application called HoaxGuard, which can be directly utilized by users. The developed system is capable of providing real-time classification results through a simple and accessible interface, making it potentially valuable as a tool for improving public digital literacy and reducing the spread of misleading information. Nevertheless, this study still contains several limitations, particularly regarding the scope of the dataset and the exclusive focus on text-based classification. Therefore, future research is recommended to employ larger and more diverse datasets while extending the system to incorporate multimodal features such as images and news metadata. In addition, the exploration of ensemble-based approaches may provide opportunities to further improve the accuracy and robustness of the proposed system in future developments.

REFERENCES

- Ahmed, H., Traore, I., & Saad, S. (2022). Detecting opinion spams and fake news using text classification. *Journal of Information Security and Applications*, 65, 103082. <https://doi.org/10.1002/spy2.9>
- Aji, A. F., Winata, G. I., Koto, F., et al. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. *Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2022.acl-long.500>
- Alnabhan, M. Q., & Branco, P. (2024). Fake news detection using deep learning: A systematic literature review. *IEEE Access*, 12, 114435–114459. <https://doi.org/10.1109/ACCESS.2024.3435497>
- Buda, M., Maki, A., & Mazurowski, M. A. (2021). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011>
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of law school. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. <https://doi.org/10.18653/v1/2020.findings-emnlp.261>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Cistia Sukmawati, E., Suryaningrum, L., Angelica, D., & Ghaniaviyanto Ramadhan, N. (2024). Klasifikasi berita palsu menggunakan model Bidirectional Encoder Representations from Transformers (BERT). <https://doi.org/10.37278/sisinfo.v6i2.934>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*. <https://doi.org/10.18653/v1/N19-1423>
- Ellam, I. V., Okorie, K. M., & Okebanama, U. F. (2025). Fake news detection system using natural language processing: An optimized approach. *European Journal of Applied Science, Engineering and Technology*, 3(2). [https://doi.org/10.59324/ejaset.2025.3\(2\).15](https://doi.org/10.59324/ejaset.2025.3(2).15)
- Ibrahim, A. A., Ali, H. U., Yakubu, I. Z., & Lawal, I. A. (2024). Fake news detection in Hausa language using transfer learning method. *International Journal of Innovative Science and Research Technology*, 9(10), 2259–2269. <https://doi.org/10.38124/ijisrt/IJISRT24OCT1050>
- Jocelynne, C., Wijayakusuma, I. G. N. L., & Harini, L. P. I. (2024). Detection of political hoax news using fine-tuning IndoBERT. *Journal of Applied Informatics and Computing*, 9(2). <https://doi.org/10.30871/jaic.v9i2.8989>
- Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80, 11765–11788. <https://doi.org/10.1007/s11042-020-10183-2>
- Khalkia, I. (2023). Analisis berita hoax dalam bahasa Indonesia menggunakan metode multilingual BERT.
- Khan, J. Y., Khondaker, M. T. I., Iqbal, A., & Afroz, S. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4, 100032. <https://doi.org/10.1016/j.mlwa.2021.100032>
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *Proceedings of COLING 2020*. <https://doi.org/10.18653/v1/2020.coling-main.66>
- Maulud, D. H., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4), 140–147. <https://doi.org/10.38094/jastt1457>
- Mujilahwati, S., Zamroni, M. R., & Sholihin, M. (2026). Hybrid deep learning approach for Indonesian hoax detection: A comparative evaluation with IndoBERT. *International Journal of Advances in Applied Sciences*, 15(1), 322–332. <https://doi.org/10.11591/ijaas.v15.i1.pp322-332>
- Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. *Proceedings*

- of COLING 2020. <https://doi.org/10.18653/v1/2020.coling-main.75>
- Praha, T., Widodo, W., & Nugraheni, M. (2024). Indonesian fake news classification using transfer learning in CNN and LSTM. *International Journal on Informatics Visualization*, 8(3). <http://dx.doi.org/10.62527/joiv.8.3.2126>
- Prayoga, H., & Kurniawan, R. (2025). Implementasi algoritma C4.5 pada klasifikasi status gizi balita menggunakan framework Flask. *Jurnal Decode*. <https://doi.org/10.51454/decode.v5i2.1194>
- Riadi, A. T., Indriani, F., Mazdadi, M. I., Faisal, M. R., & Herteno, R. (2025). Cross-temporal generalization of IndoBERT for Indonesian hoax news classification. *Jurnal Teknik Informatika (JUTIF)*. <https://doi.org/10.52436/1.jutif.2025.6.5.4757>
- 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*, 10(3). <https://doi.org/10.26555/jiteki.v10i3.29450>
- Rizqullah, M. R., et al. (2023). Indonesian fact and hoax political news [Data set]. Kaggle. <https://www.kaggle.com/datasets/linkgish/indonesian-fact-and-hoax-political-news>
- Roy, A., Basak, K., Ekbal, A., & Bhattacharyya, P. (2022). A deep ensemble framework for fake news detection using transformer-based models. *Expert Systems with Applications*, 198, 116758. <https://doi.org/10.48550/arXiv.1811.04670>
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 160. <https://doi.org/10.1007/s42979-021-00592-x>
- Sholikhah, I. I., Tri, A., Harjanta, J., & Latifah, K. (2023). Machine learning untuk deteksi berita hoax menggunakan BERT. *Proceedings of INFEST 2023*, 1–8. <https://conference.upgris.ac.id/index.php/infest/article/view/3818>
- Shu, K., Wang, S., & Liu, H. (2020). Beyond news contents: The role of social context for fake news detection. *Proceedings of WSDM*. <https://doi.org/10.48550/arXiv.1712.07709>
- Suherlan, E., Arti, S., Nabilah, S., & Hazimah, Z. (2025). Penerapan Flask framework untuk deployment model machine learning dalam mendukung analisis adaptasi mahasiswa pada pembelajaran daring. *Jurnal PINTER*. <https://doi.org/10.21009/pinter.9.1.15>
- Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2021). EANN: Event adversarial neural networks for multi-modal fake news detection. *Proceedings of the ACM SIGKDD*. <https://doi.org/10.1145/3219819.3219903>
- Wibawa, I. G. B. S., Indrayana, I. N. E., & Ariawan, M. P. A. (2024). Penerapan metode IndoBERT untuk deteksi berita hoaks pada media digital berbahasa Indonesia. *Jurnal Riset dan Aplikasi Mahasiswa Informatika (JRAMI)*. <https://doi.org/10.30998/94zt3k10>
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S. Y., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *Proceedings of EMNLP 2020*.