Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

Cross-Temporal Generalization of IndoBERT for Indonesian Hoax News Classification

Agus Teguh Riadi*[1], Fatma Indriani[2], Muhammad Itqan Mazdadi[3], Mohammad Reza Faisal[4], Rudi Herteno[5]

[1,2,3,4,5]Computer Science Department, Lambung Mangkurat University, Indonesia

Email: [1]f.indriani@ulm.ac.id

Received : May 20, 2025; Revised : Oct 27, 2025; Accepted : Oct 29, 2025; Published : Oct 31, 2025

Abstract

The spread of hoaxes in digital media poses a major challenge for automated detection systems as language and topics evolve over time. Although Transformer-based models such as IndoBERT have demonstrated high accuracy in previous studies, their performance across different time periods remains underexplored. This study examines the cross-temporal generalization ability of IndoBERT for hoax news classification. The model was trained on labeled articles from 2018–2023 and tested on data from 2025 to evaluate its robustness against temporal distribution shifts. The results indicate high accuracy on similar-period data (99.67–99.89%) but a decrease on 2025 data (95.45–95.87%), with most errors occurring as false negatives in the hoax class. These findings highlight the impact of temporal distribution shifts on model reliability and underscore the importance of adaptive strategies such as periodic retraining and domain-based data augmentation. Practically, this model has the potential to assist social media platforms and government institutions in developing dynamic and time-adaptive hoax detection systems. The cross-temporal approach employed in this study also offers methodological innovation compared to conventional random validation, as it better reflects real-world conditions where misinformation patterns continually evolve.

Keywords : *Cross-set, Hoax Detection, IndoBERT, Model Generalization, Temporal Distribution Shift.*

1. INTRODUCTION

The rapid advancement of Internet technology and social media over the past few decades has fundamentally changed how people access and share information. With easy access and near-instantaneous dissemination, information can now reach millions of people within seconds[1]. However, this convenience also brings serious challenges, particularly the widespread circulation of false information, or hoaxes[2]. A hoax refers to information that is deliberately created and distributed with the intent to mislead readers or the broader public[3]. Such content often exhibits identifiable linguistic characteristics, including disjointed or overly simplistic sentence structure, sensational or emotionally charged language, and cherry-picked "facts" that lack credible sources[4]. In the absence of editorial oversight, hoaxes can spread rapidly, taking advantage of cognitive biases such as confirmation bias and emotional contagion to encourage sharing[5].

The negative effects of hoax dissemination are both immediate and far-reaching. Politically motivated hoaxes can erode public trust in elections and democratic institutions, fueling polarization and civil unrest[6]. In public health contexts, misinformation about vaccines or treatments can lower compliance with medical guidance, leading to preventable illness and increased mortality[7]. During emergencies or natural disasters, hoaxes may disrupt crisis response efforts by spreading false alarms or diverting resources[8]. In short, the unchecked spread of hoaxes threatens public safety, undermines institutional credibility, and damages the overall integrity of the information ecosystem[9].

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

To address these risks, researchers have turned to automated hoax detection systems, leveraging recent advances in natural language processing (NLP) and machine learning[10]. Early approaches ranged from classical machine learning methods such as Support Vector Machines (SVM) and Random Forests using TF-IDF features, to sequence-based models like LSTM. However, Transformer-based models have shown superior performance across a range of NLP tasks[11][12]. Among these, BERT (Bidirectional Encoder Representations from Transformers) offers key advantages due to its ability to capture bidirectional context and build rich semantic representations using self-attention mechanisms[13]. Unlike traditional models that rely on local word patterns or frequency-based features, BERT can model long-range dependencies and identify subtle inconsistencies and emotional cues often present in hoax content[14][15].

Several recent studies have applied BERT and its variants to hoax detection in the Indonesian language. Awalina et al.[16] fine-tuned a base BERT model on a dataset of 1,116 articles from TurnBackHoax.id, achieving 90% accuracy—outperforming CNN (85%) and BiLSTM (87%). Sinapoy et al.[17] compared IndoBERT and LSTM for hoax classification on Twitter text, where IndoBERT reached 92.07% accuracy, surpassing LSTM's 87.54%. Suadaa et al.[18] also found IndoBERT to be more effective than BERT and mBERT when trained on a COVID-19 hoax corpus, achieving 97.7% accuracy. These findings support the use of Indonesian-specific pre-trained models for tasks involving local language content.

However, most of these studies evaluate models under the assumption that training and test data are drawn from the same distribution. In practice, hoaxes evolve rapidly in topic, vocabulary, and style[19]. Terms or political references that are prominent one year may fade the next, while new terminology can emerge quickly in response to current events. This kind of distributional change (sometimes referred to as *concept drift*) can reduce model performance when the patterns learned during training no longer match those found in new data[20]. One specific form of concept drift is *vocabulary drift*, where the distribution of keywords or subword tokens shifts over time, leading to weaker representations of unseen or rare terms[21]. Additionally, variation in text length can also affect performance, as short texts may not provide sufficient context for reliable classification[22]. Recent work has shown that temporal drift can significantly degrade model performance. Zhao et al.[23] demonstrated that language evolution over time leads to reduced accuracy when models are applied to newer test data. Chalkidis et al.[24] further found that performance drops more sharply when models are tested on chronologically split datasets compared to randomly split ones, underscoring the real-world impact of temporal shifts.

Despite this, few studies have examined IndoBERT's generalization ability across different time periods. Most rely on random data splits, which fail to reflect real-world conditions where future content often differs from past data. This study addresses that gap by systematically evaluating IndoBERT's temporal generalization through a cross-set approach, training on Indonesian news and hoax data from 2020–2023 and testing on newer data from 2025. This evaluation approach assumes that the chronological split of data (2020–2023 for training vs. 2025 for testing) serves as a valid representation to simulate real-world temporal concept drift challenges. In contrast to previous studies that focus solely on in-distribution performance, this work compares intra-set and cross-set results to assess IndoBERT's robustness to evolving language and topic patterns.

This study makes several key contributions. Firstly, it evaluates IndoBERT's ability to generalize across time in the context of Indonesian hoax detection, filling a gap in prior research that has not systematically addressed temporal drift. Secondly, it analyzes the impact of changes in text length, vocabulary, and topical focus on model performance, providing insights into the challenges of deploying NLP models in real-world, time-sensitive applications. Therefore, it offers concise practical

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

recommendations for government and social media platforms, such as regularly updating models with current data to combat the continuously evolving spread of false information.

2. METHOD

This research aims to develop and evaluate a classification model for detecting hoax news in Indonesia using IndoBERT, with a specific focus on assessing the model's ability to generalize across time. The process begins with data collection from two sources: a Kaggle dataset containing historical news from 2018–2023 and web-crawled articles from CNBC Indonesia and TurnBackHoax representing 2025 content. This combination simulates real-world deployment scenarios where models trained on past data must classify newer hoaxes with evolving vocabulary, topics, and writing styles. The implementation is conducted in Google Colab using PyTorch and Hugging Face Transformers, supported by Pandas, NumPy, and Scikit-learn for data processing and evaluation

Each dataset is preprocessed to clean and normalize the text by removing metadata, special characters, and irrelevant tokens. The text is then tokenized with [CLS], [SEP], and [PAD] tokens to match the BERT input format. The older dataset is divided into training and validation sets to fine-tune IndoBERT for binary classification (hoax vs. factual). Evaluation is performed on both in-distribution and out-of-distribution data to assess temporal generalization. Model performance is measured using accuracy, precision, recall, and F1-score. Figure 1 presents the overall workflow, covering preprocessing, tokenization, model fine-tuning, and comparison between intra-set and cross-set performance.

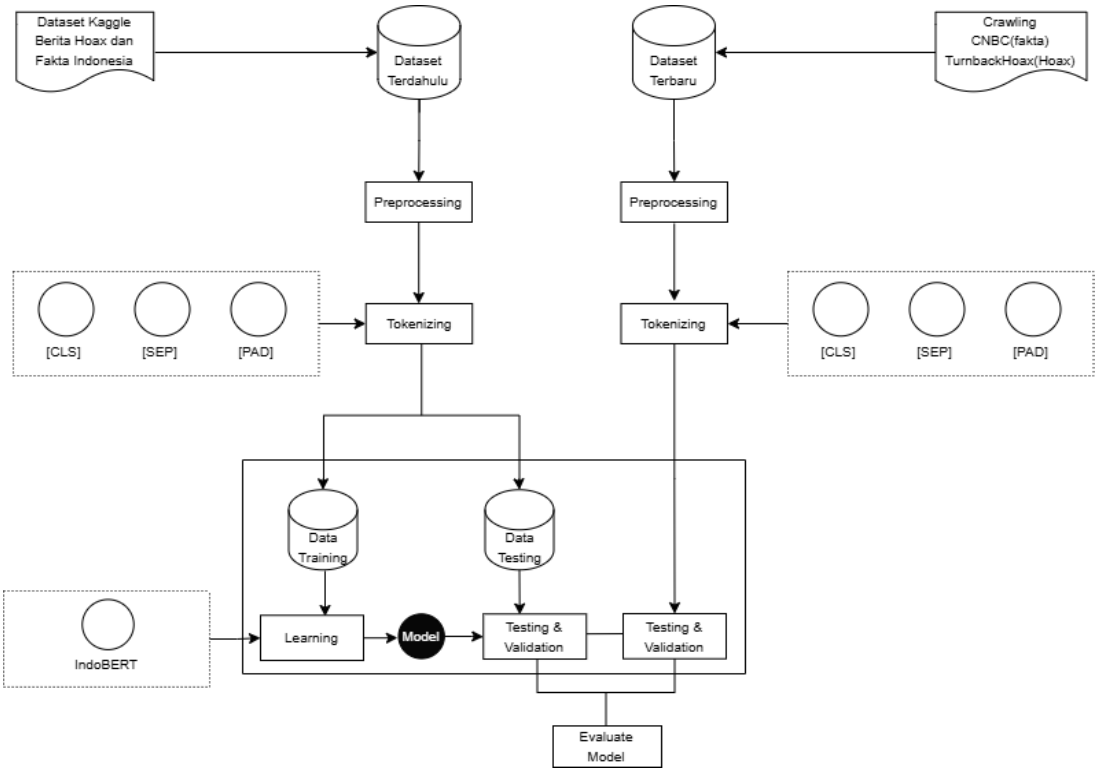

Figure 1. Research flowchart

2.1. Dataset

In this study, the training data were sourced from Kaggle and comprise two categories: factual news from CNBC Indonesia and hoax news from the TurnBackHoax. The testing data were collected via automated web scraping, including the latest factual news from CNBC Indonesia and recent hoax

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

reports from TurnBackHoax. The data collection period spans different years to ensure a diverse range of topics and capture the evolution of news writing styles. The composition of the data used in this study is shown in Table 1.

Table 1. Dataset Composition

Source	News Type	News Period	Volume
Kaggle (CNBC Indonesia)	Factual	2023	4505
Kaggle (TurnBackHoax/Mafindo)	Hoax	2018-2023	4590
Web Scraping (CNBC Indonesia)	Factual	2025	4000
Web Scraping (TurnBackHoax)	Hoax	2023-2025	3987

2.2. Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a deep language model built on the Transformer encoder architecture[25]. In its base configuration, BERT consists of 12 stacked transformer layers. Each layer contains multi-head self-attention and feed-forward sublayers[26].

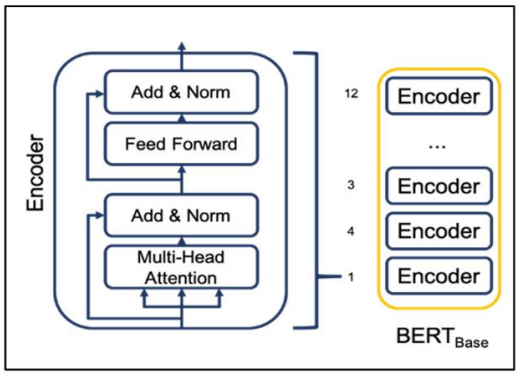

Figure 2. BERT Architecture [27]

The first sub-layer is the multi-head self-attention mechanism, which is the process of running several self-attentions in parallel with different sets of weights. The equation of self-attentions can be seen in (1), while the Multi-head self-attention equation can be seen in (2), (3)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK}{\sqrt{d_k}}\right)V \quad (1)$$

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

The second sub-layer is the feed-forward layer, which helps the model capture non-linear relationships and refine the information processed by the self-attention mechanism. The equation of feed-forward neural can be seen in (4)

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

This mechanism enabling the model to compute contextualized representations that consider all tokens in an input sequence simultaneously. In other words, BERT's attention is bidirectional: every token's representation is informed by the full left and right context[28]. This bidirectional training (via

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

masking) allows BERT to capture rich context from both directions, leading to superior language understanding.

2.3. BERT Text Representation

Before BERT processes text data, the data is first converted into the model's required input format through several steps: tokenization, padding, numericalization, and embedding[29]. BERT splits the input text into tokens using the WordPiece model, then adds the special token [CLS] at the beginning and [SEP] at the end. Before further processing, each token sequence within a batch is length-aligned through padding that is, shorter sequences are padded with a special token (e.g., [PAD]) so that all sequences in the batch share the same length[30]. Each sub-word token or BERT special token (e.g., [CLS], [SEP], [PAD) is mapped to an integer indicating its position in the vocabulary list[31]. The final representation of each token is formed by summing three types of embeddings[32], as shown in Figure 3, token embeddings, which capture the token's semantic meaning; segment embeddings, which distinguish between sentence segments in the input; and position embeddings, which encode the token's position in the sequence. This summed vector is then used as input to the self-attention layers within the Transformer architecture.

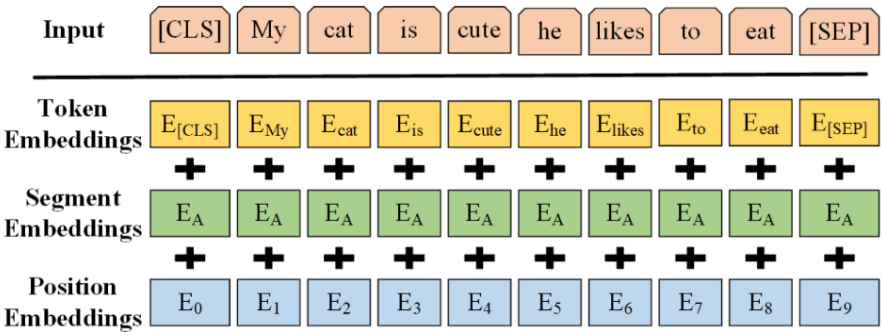

Figure 3. BERT Input Representation[33]

2.4. Indonesia Based on BERT (IndoBERT)

IndoBERT is a pre-trained language model based on the BERT architecture, specifically developed for the Indonesian language. This model follows the standard BERT-Base configuration with 12 transformer encoder layers, a hidden dimension size of 768, 12 self-attention heads, and a feed-forward dimension of 3072. IndoBERT is trained using the Masked Language Model task (without Next Sentence Prediction) on a large Indonesian corpus (approximately 220 million words from Wikipedia, news, web, etc.). The model's vocabulary uses WordPiece tokens with around 31,923 tokens[34]. As a result, IndoBERT can capture contextual representations that align with the unique structure and vocabulary of the Indonesian language. The choice of IndoBERT is based on its effectiveness in handling Indonesian. Due to its pre-training on an Indonesian corpus, this model is more accurate in understanding local context compared to general multilingual models[35].

2.5. Cross-set Generalization

Cross-set generalization is the ability of a model to produce accurate predictions on data that has never been seen before, especially when the data comes from a different set or dataset than the one used during training[36]. This concept becomes especially important when the test data come from a dataset different from the original training data. A model with good generalization not only learns specific patterns from the training set, but also acquires the essential features or structures that can be applied to

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

new datasets[37]. In the context of this research, cross-set evaluation is conducted by training the model on old training data and then testing it on new data, thereby assessing how well the model can recognize hoax patterns that have never been encountered before, so that its performance can be evaluated in dynamic real-world situations.

2.6. Experimental Setup

2.6.1. Data Preprocessing

The TurnBackHoax dataset consists of complete narratives resulting from official fact-checking of hoax claims along with their original context; because the model's objective is to identify only the core false claim, each "Full Narrative" is first condensed into a "Hoax Narrative" that preserves the main statement. Similarly, the CNBC dataset which includes the headline, byline, publication date, and full article body is filtered into a "Clean Article" containing only the headline and essential article text. The extracted content from both datasets then undergoes corpus normalization, removing HTML tags are stripped, editorial annotations in square brackets are excised, "@" handles and hyperlinks are removed, and all emojis or non-verbal Unicode symbols are eliminated to ensure consistent input formatting for downstream embedding models

2.6.2. Training Data Proportions

Variations in training data proportions of 70%, 80%, and 90% were used to measure the impact of data volume on model performance: first, in experiments 1–3, the model was trained and tested intra-set (old→old) with training–testing ratios of 70:30 (small), 80:20 (medium), and 90:10 (large) as the baseline; then, in experiments 4–6, the model was trained on old data with the same proportions but tested on new data (cross-set old→new) to assess generalization at small, medium, and large training scales.

2.6.3. Fine-Tuning IndoBERT Model

The standard strategy for leveraging pre-trained models like BERT is to fine-tune them so they can be adapted to a specific task[38]. Fine-tuning is the process of adapting a pretrained BERT model to perform optimally on a specific task by adjusting its weights based on new, relevant data. During fine-tuning, the model is taught to pay special attention to tokens like [CLS], which are used for classification, and to disregard tokens like [SEP], which serve merely as separators[26].

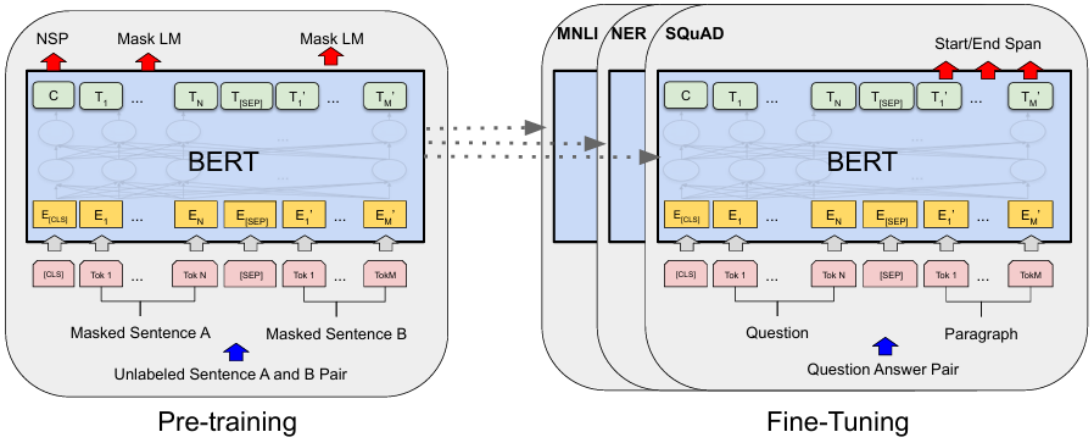

Figure 4. BERT Pre-training and Fine-tuning Phases[39]

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

In this study, we use the pre-trained indoBERT model from "indolem/indobert-base-uncased" model. Tokenization is performed with `AutoTokenizer`, which adds the special tokens `[CLS]` and `[SEP]`, applies truncation (up to 512 tokens), and pads sequences to produce input IDs and attention masks. Training is optimized using AdamW with an initial learning rate of 3×10^{-5} and an epsilon of 1e-8. The training was run up to 5 epochs with early stopping monitoring the validation loss; evaluation was performed each epoch and the checkpoint with the lowest validation loss was saved automatically. Training stopped if no improvement occurred within a patience of 1–2 epochs. The use of a learning-rate scheduler, early stopping, and best-checkpoint selection aims to stabilize training and improve generalization to out-of-distribution (OOD) data.

2.6.4. Performance Evaluation

After the training phase, the model is tested in a cross-dataset scenario using new data and its performance evaluated for detecting hoax patterns beyond the scope of the training set. In classification model evaluation, selecting appropriate metrics is crucial to avoid bias[40]. The confusion matrix is commonly used because it effectively displays the comparison between the model's predictions and the ground truth[41]. In a confusion matrix, there are several important terms such as True Positive (TP): the model predicts positive and the actual result is also positive; True Negative (TN): the model predicts negative and the actual result is also negative; False Positive (FP): the model predicts positive, but the actual result is negative; and False Negative (FN): the model predicts negative, but the actual result is positive. The terms are defined in Table 2.

Table 2. Confusion Matrix

Actual Class	**Predicted Class**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The performance of each component is assessed using an evaluation matrix that incorporates the elements of the confusion matrix.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \times \mathbf{100\%} \quad (5)$$

$$\boldsymbol{F1-Score} = \mathbf{2} \times \frac{\boldsymbol{Recall \times Precision}}{\boldsymbol{Recall+Precision}} \times \mathbf{100\%} \quad (6)$$

$$Precision = \frac{TP}{FP+TP} \times 100\% \quad (7)$$

$$Recall = \frac{TP}{FN+TP} \times 100\% \quad (8)$$

3. RESULT

3.1. Intra Set versus Cross Set Performance

This section presents the empirical results of the classification experiments summarized in Table 3. For each train–test split configuration (70:30, 80:20, and 90:10), the model's accuracy on intra-set evaluation (old→old) remained consistently high, ranging from 0.9967 to 0.9989. This confirms that IndoBERT performs well when tested on data from the same temporal distribution as its training data.

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

In contrast, accuracy declined to between 0.9545 and 0.9587 in the cross-set evaluation (old→new), indicating that the model had more difficulty classifying texts from a different time period (2025).

Table 3. Accuracy and Confusion-Matrix Breakdown Intra-set vs. Cross-set Evaluation

Training Split	Test Type	Accuracy	(TP)	(FP)	(TN)	(FN)
70 : 30	Intra	0.997	1357	1	1366	5
70 : 30	Cross	0.954	1082	13	1206	96
80 : 20	Intra	0.998	918	0	899	2
80 : 20	Cross	0.956	752	43	777	26
90 : 10	Intra	0.996	442	0	465	3
90 : 10	Cross	0.958	360	5	406	28

Most of the errors in the intra-set tests were false negatives in the hoax class. For instance, there were 5 such errors in the 70:30 split, 2 in the 80:20 split, and 3 in the 90:10 split. In the cross-set setting, the number of errors increased substantially: 109 total misclassifications with the 70:30 split and 33 with the 90:10 split. These were again dominated by false negatives in the hoax class—96 and 28 instances, respectively. Interestingly, in the 80:20 split, the dominant error type shifted to false positives in the factual class (43 cases). The results indicate that increasing the training size generally improves cross-set accuracy, though the gains are not linear. For example, increasing training data from 70% to 90% improved cross-set accuracy by only 0.4 percentage points. Although the overall drop in accuracy appears small (less than 5%), in practice, this reduction may lead to a substantial number of hoaxes going undetected, especially when scaled to high-volume news or social media feeds. Therefore, even small shifts in performance should be interpreted carefully, particularly in real-world applications where undetected misinformation can cause public harm. This motivates the need for adaptive modeling approaches that can handle temporal variation in both vocabulary and narrative structure.

3.2. Distributional Shift

This section examines two main factors contributing to IndoBERT's reduced performance on newer data: differences in text length (Figure 4) and vocabulary drift (Figure 5) between the training and test corpora. First, token-length distributions show a clear shift. During training, most CNBC articles ranged from 200 to 600 tokens, peaking at around 400. In contrast, the 2025 CNBC test articles tended to be shorter (peaking near 300 tokens), with some very long documents reaching 2,500 tokens that required truncation. Excessive truncation led to the loss of important information at the end of texts[42], Meanwhile, very short texts receiving heavy padding has been reported to degrade the quality of their representations and therefore damages the classification accuracy [43]. This finding is in line with the work Søgaard's et al.[24] observations that the passage of time between training and test data affects performance most among several other factors including the length of text.

Second, vocabulary drift limits the model's ability to represent unfamiliar terms. As shown in Figure 5, 21,879 subword tokens are shared between training and test sets, but 4,483 are exclusive to the training data, and 992 are new in the test set. These new tokens reflect changes in public discourse and terminology that IndoBERT did not encounter during fine-tuning. Prior work[23] has shown that such shifts like the emergence of new pandemic-era terms can reduce model reliability on recent content. Since embeddings for unseen tokens are less informative, the model's understanding of new text is weakened.

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

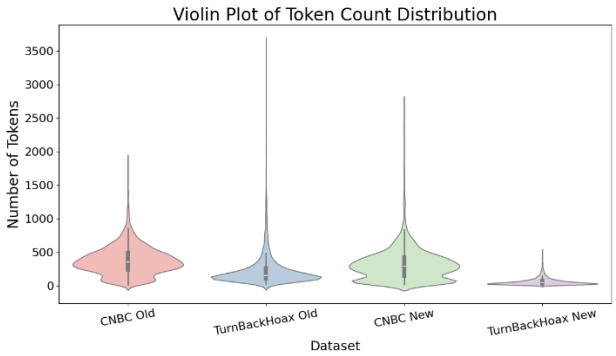

Figure 4. Token Count Distribution Across Subsets

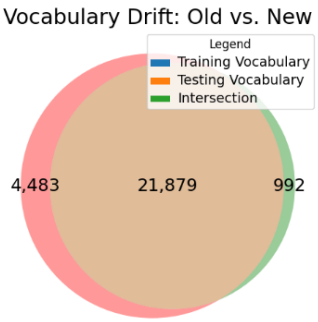

Figure 5. Vocabulary Drift old vs new data

These shifts in input length and vocabulary illustrate how IndoBERT's strong performance on familiar data does not fully carry over to newer content. As further shown in Table 4, the most frequent tokens differ between old and new subsets, indicating not only a change in vocabulary, but also in dominant topics and writing style. This reinforces the need for continuous updates or adaptation mechanisms when deploying static language models in dynamic environments.

Table 4. Top 5 Most Frequent Tokens (with Counts) in Old vs. New Subsets

CNBC_Old	CNBC_New	TurnbackHoax_Old	TurnbackHoax_New
rp (9797)	indonesia (7696)	the (4008)	indonesia (1037)
indonesia (8520)	tahun (5458)	indonesia (3276)	prabowo (797)
tahun (6673)	rp (4723)	tersebut (3269)	jokowi (668)
tersebut (6384)	tersebut (4204)	orang (3095)	ani (553)
bank (6295)	menjadi (3824)	foto (2471)	presiden (538)

4. DISCUSSION

4.1. Factors Affecting Cross-Temporal Performance

The performance gap between intra-set and cross-set evaluations observed in Table 3 suggests that IndoBERT's accuracy is sensitive to temporal distribution shifts. While the model achieved high accuracy (above 99.6%) when tested on data from the same time period as its training set, accuracy declined to the 95.45–95.87% range when applied to newer data from 2025. This drop, though numerically small, represents a meaningful degradation in real-world scenarios, particularly in high-volume environments where small misclassification rates can lead to significant numbers of undetected hoaxes. A deeper examination of the dataset reveals two main factors that likely contributed to this decline: (1) changes in input length, and (2) vocabulary drift.

Figure 4 shows a distinct change of token-lengths distribution, where training and test data are used. The training data, including that from CNBC, was heavily centered around medium length articles (with a peak of 400 tokens). Compared to 2025, which has much shorter texts on average (though there are some articles longer than 2500 tokens) with a peak at around 300 tokens. Both extremes introduce classification challenges. Short strings lack sufficient context to avoid misclassification, while long ones are potentially truncated during processing because of the limited model input length. This truncation causes the loss of valuable information towards the end of longer articles resulting in a lower classification accuracy. Vocabulary shift also contributes to the drop in performance. As shown in Figure 5, while most of the subword tokens are also shared between old vs. new datasets, test set from 2025

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

contributes a total of 992 never-before-seen tokens. These new tokens indicate shifts in topical concentration or linguistic usage (e.g., political referents or culturally-specific allusions developed after the training period). Without prior exposure to these terms during fine-tuning, IndoBERT's embedding space may lack the representational depth needed to fully interpret the new input.

The top-token frequencies in Table 4 provide additional evidence for these temporal shifts. The most common terms from each sub-corpus (Table 4) imply fluctuations in content emphasis within time periods. In the old CNBC data, terms like rp (currency) and bank occur more frequently, suggesting that it is full of financial news. The more recent article topics tend to be less specific, with keywords include tahun (year), indonesia and menjadi (i.e., to become), indicating general topic reporting. The TurnBackHoax dataset also exhibits the same trend: in old entries, we only find generic words such as tersebut (that) and foto (photo), whereas its younger range consists of more named entities like prabowo, jokowi, and ani that are relevant to political matters that come up around 2024–2025. These shifts in vocabulary and topical references may contribute to IndoBERT's reduced performance on the newer data. Together, these findings highlight the difficulty of applying a static, fine-tuned model like IndoBERT to evolving real-world data. Even modest changes in language patterns, article structure, and terminology can noticeably reduce classification performance if the model is not regularly updated to reflect current usage.

4.2. Qualitative Error Analysis

Table 5 presents examples of misclassified news snippets. In the first two cases, factual reports were incorrectly labeled as hoaxes. In the last two, hoax content was incorrectly predicted as factual. These types of errors indicate challenges in the model's ability to generalize to unseen language patterns, particularly when phrasing or context diverges from the training data.

Table 5. Example of Misclassification on Test Data

News(Snippet)	Actual Label	Predict Label
presiden prabowo subianto berkeinginan untuk mensejahterakan petani di Indonesia…(*President Prabowo Subianto wants to improve the welfare of farmers in Indonesia…*)	Fact	Hoax
Sebanyak 31 orang tewas dilaporkan tewas setelah sebuah bus penumpang terjun…(*A total of 31 people were reported dead after a passenger bus plunged…*)	Fact	Hoax
baru saja gempa magnitude 10, 6 guncang maluku hingga hancur terbelah…(*Just now a magnitude 10.6 earthquake shook Maluku until it was torn apart…*)	Hoax	Fact
jokowi restui perpanjangan izin tambang freeport seumur cadangan tapi izin freeport …(*Jokowi approves extension of Freeport's mining permit for the lifetime of the reserve but Freeport's permit…*)	Hoax	Fact

One potential cause is the model's reliance on surface-level token associations. Prior studies have shown that BERT-based classifiers are susceptible to spurious correlations, where frequently co-occurring tokens are overemphasized during training[44]. For instance, political names such as *Prabowo* or *Jokowi* (rows 1 and 4) may be associated with hoax narratives in the training set, leading the model to misclassify new content mentioning them—even when the content is factual. Additionally, semantic ambiguity or exaggerated phrasing can confuse the model. The earthquake headline in row 3 describes an implausibly large event, while the bus accident in row 2 uses emotionally charged language. Without additional context, these may resemble hoaxes, resulting in incorrect predictions[45]. These cases highlight the limits of relying solely on token-level semantics, and point to the potential benefits of

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

incorporating external knowledge or contextual signals to improve classification accuracy in dynamic news environments.

4.3. Limitation and Future Direction

This study has several limitations, primarily related to the static training setup and the evolving nature of language in online media. The IndoBERT model was fine-tuned only once using historical data from two fixed sources, CNBC Indonesia and TurnBackHoax, and was not updated thereafter. As a result, the model is unable to adapt to domain or temporal shifts, such as the emergence of new vocabulary or topics in the 2025 data. This limitation contributes to the observed decline in classification performance when tested on newer content. Moreover, the dataset is restricted to two online news platforms, which may not fully capture the range and variety of hoaxes circulating across different media channels or social platforms. This limited domain coverage affects the model's generalization ability in more diverse real-world scenarios.

To address these issues, future work should consider affordable and incremental adaptation strategies. One practical approach is scheduled retraining, where the model is periodically fine-tuned on newly collected and labeled data to stay aligned with evolving language use. Previous studies have shown that this method can significantly improve out-of-time performance[46]. In addition, lightweight text augmentation techniques can be applied to increase training diversity without requiring new annotations. Examples include modifying text length, deleting or substituting non-essential words, or paraphrasing through methods such as back-translation or EDA[47]. In this way, the study not only provides an empirical overview of IndoBERT's performance in intra-set and cross-set scenarios, but also highlights the main challenges that must be addressed to build a hoax detection system that remains effective in the face of continuous linguistic and topical change.

5. CONCLUSION

This study evaluated IndoBERT's performance under temporal distribution shifts in Indonesian hoax detection by comparing intra-set and cross-set results across 70:30, 80:20, and 90:10 training–test splits. While intra-set accuracy remained consistently high (99.67–99.89%), cross-set accuracy declined to 95.45–95.87%, reflecting a performance drop of less than 5% when applied to 2025 data. Further analysis suggested that shifts in text length, vocabulary, and topical focus contributed to this performance gap. The model struggled with content that differed linguistically or semantically from the training data, particularly when newer topics or named entities appeared. These findings confirm that static fine-tuning is insufficient for maintaining performance over time.

To enhance adaptability, future work should integrate periodic retraining with updated corpora, text augmentation to simulate linguistic variation, and approaches such as continual learning or domain adaptation to address evolving vocabularies and topics. Implementing these techniques could enable social media or government platforms to detect hoaxes more adaptively, ensuring timely and context-aware identification of misinformation as online discourse evolves.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, p. 30, 2023, doi: 10.1007/s13278-023-01028-5.

[2] E. Denniss and R. Lindberg, "Social media and the spread of misinformation: infectious and a

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

threat to public health," *Health Promot. Int.*, vol. 40, no. 2, 2025, doi: 10.1093/heapro/daaf023.
[3] Maria D Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee, "'Fake News' Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content," *Am. Behav. Sci.*, vol. 65, no. 2, pp. 180–212, Oct. 2019, doi: 10.1177/0002764219878224.
[4] C. Zhou, K. Li, and Y. Lu, "Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness," *Inf. Process. Manag.*, vol. 58, no. 6, p. 102679, 2021, doi: 10.1016/j.ipm.2021.102679.
[5] S. Chen, L. Xiao, and A. Kumar, "Spread of misinformation on social media: What contributes to it and how to combat it," *Comput. Human Behav.*, vol. 141, p. 107643, 2023, doi: https://doi.org/10.1016/j.chb.2022.107643.
[6] A. Wibawa, H. Joko Utomo, A. Bagus, and I. Dwi Arianto, "Political Hoaxes on Social Media Before the 2024 Regional Elections," *SHS Web Conf.*, vol. 212, no. August 2018, p. 04047, 2025, doi: 10.1051/shsconf/202521204047.
[7] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA," *Nat. Hum. Behav.*, vol. 5, no. 3, pp. 337–348, 2021, doi: 10.1038/s41562-021-01056-1.
[8] S. Torpan *et al.*, "Handling false information in emergency management: A cross-national comparative study of European practices," *Int. J. Disaster Risk Reduct.*, vol. 57, p. 102151, 2021, doi: 10.1016/j.ijdrr.2021.102151.
[9] S. Muhammed T and S. K. Mathew, "The disaster of misinformation: a review of research in social media.," *Int. J. data Sci. Anal.*, vol. 13, no. 4, pp. 271–285, 2022, doi: 10.1007/s41060-022-00311-6.
[10] K. D. Patel, "Fake News Detection on Natural Language Processing A Survey," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 9, pp. 115–121, 2020, doi: 10.26438/ijcse/v7i9.115121.
[11] Valentina Porcu, "Past vs. Present: Key Differences Between Conventional Machine Learning and Transformer Architectures," *Adv. Nonlinear Var. Inequalities*, vol. 28, no. 2s, pp. 244–262, 2024, doi: 10.52783/anvi.v28.2537.
[12] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Caleanu, "Natural language processing with transformers: a review," *PeerJ Comput. Sci.*, vol. 10, pp. 1–22, 2024, doi: 10.7717/PEERJ-CS.2222.
[13] A. Fatwanto, F. Zamakhsyari, and R. Ndungi, "A Systematic Literature Review of BERT-based Models for Natural Language Processing Tasks," *J. Infotel*, vol. 16, no. 4, pp. 713–728, 2024, doi: 10.20895/infotel.v16i4.1206.
[14] R. Anggrainingsih, G. M. Hassan, and A. Datta, "Evaluating BERT-based language models for detecting misinformation," *Neural Comput. Appl.*, vol. 37, no. 16, pp. 9937–9968, 2025, doi: 10.1007/s00521-025-11101-z.
[15] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021, doi: 10.1007/s11042-020-10183-2.
[16] J. Fawaid, A. Awalina, R. Y. Krisnabayu, and N. Yudistira, "Indonesia's Fake News Detection using Transformer Network," in *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, 2021, pp. 247–251. doi: 10.1145/3479645.3479666.
[17] Muhammad Ikram Kaer Sinapoy, Yuliant Sibaroni, and Sri Suryani Prasetyowati, "Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 3, pp. 657–662, 2023, doi: 10.29207/resti.v7i3.4830.
[18] L. H. Suadaa, I. Santoso, and A. T. B. Panjaitan, "Transfer Learning of Pre-trained Transformers for Covid-19 Hoax Detection in Indonesian Language," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 3, p. 317, 2021, doi: 10.22146/ijccs.66205.
[19] W. Ansar and S. Goswami, "Combating the menace: A survey on characterization and detection of fake news from a data science perspective," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100052, 2021, doi: 10.1016/j.jjimei.2021.100052.
[20] M. A. Hashmani, S. M. Jameel, M. Rehman, and A. Inoue, "Concept Drift Evolution In Machine Learning Approaches: A Systematic Literature Review," *Int. J. Smart Sens. Intell. Syst.*, vol. 13, no. 1, pp. 1–16, 2020, doi: 10.21307/ijssis-2020-029.

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

[21] T. A. Chang *et al.*, "Characterizing and Measuring Linguistic Dataset Drift," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 8953–8967, 2023, doi: 10.18653/v1/2023.acl-long.498.
[22] H. Wang, G. Luo, and R. Li, "A Short Text Classification Method Based on Combining Label Information and Self-attention Graph Convolutional Neural Network," *Commun. Comput. Inf. Sci.*, vol. 1330 CCIS, no. 1, pp. 670–677, 2021, doi: 10.1007/978-981-16-2540-4_50.
[23] Z. Zhao, G. Chrysostomou, K. Bontcheva, and N. Aletras, "On the Impact of Temporal Concept Drift on Model Explanations," *Find. Assoc. Comput. Linguist. EMNLP 2022*, pp. 4068–4083, 2022, doi: 10.18653/v1/2022.findings-emnlp.298.
[24] I. Chalkidis and A. Søgaard, "Improved Multi-label Classification under Temporal Concept Drift: Rethinking Group-Robust Algorithms in a Label-Wise Setting," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 2441–2454, 2022, doi: 10.18653/v1/2022.findings-acl.192.
[25] E. Shem-Tov, M. Sipper, and A. Elyasaf, "BERT Mutation: Deep Transformer Model for Masked Uniform Mutation in Genetic Programming," *Mathematics*, vol. 13, no. 5, pp. 1–17, 2025, doi: 10.3390/math13050779.
[26] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.
[27] M. H. Y. Binhammad, A. Othman, L. Abuljadayel, H. Al Mheiri, M. Alkaabi, and M. Almarri, "Investigating Advanced Generative Dialogue Systems for Educational Chatbots," *Creat. Educ.*, vol. 15, no. 08, pp. 1593–1626, 2024, doi: 10.4236/ce.2024.158096.
[28] E. Cesario, C. Comito, and E. Zumpano, "A survey of the recent trends in deep learning for literature based discovery in the biomedical domain," *Neurocomputing*, vol. 568, no. November 2023, 2024, doi: 10.1016/j.neucom.2023.127079.
[29] N. J. Prottasha *et al.*, "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning," *Sensors*, vol. 22, no. 11, pp. 1–19, 2022, doi: 10.3390/s22114157.
[30] L. M. Pham and H. C. The, "LNLF-BERT: Transformer for Long Document Classification with Multiple Attention Levels," *IEEE Access*, vol. 12, no. November, pp. 165348–165358, 2024, doi: 10.1109/ACCESS.2024.3492102.
[31] C. Toraman, E. H. Yilmaz, F. Aahi˙nuç, and O. Ozcelik, "Impact of Tokenization on Language Models: An Analysis for Turkish," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 4, 2023, doi: 10.1145/3578707.
[32] L. Stankevičius and M. Lukoševičius, "Extracting Sentence Embeddings from Pretrained Transformer Models," *Applied Sciences*, vol. 14, no. 19. 2024. doi: 10.3390/app14198887.
[33] X. Chen, P. Cong, and S. Lv, "A Long-Text Classification Method of Chinese News Based on BERT and CNN," *IEEE Access*, vol. 10, pp. 34046–34057, 2022, doi: 10.1109/ACCESS.2022.3162614.
[34] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.
[35] R. Bagestra, A. Misbullah, Z. Zulfan, R. Rasudin, L. Farsiah, and S. Azizah, "Performance Assessment of Machine Learning and Transformer Models for Indonesian Multi-Label Hate Speech Detection," vol. 2, no. 2, pp. 62–71, 2024, doi: 10.60084/ijds.v2i2.235.
[36] G. Blanchard, A. A. Deshmukh, Ü. Dogan, G. Lee, and C. Scott, "Domain generalization by marginal transfer learning," *J. Mach. Learn. Res.*, vol. 22, no. 1, Jan. 2021, doi: 10.5555/3546258.3546260.
[37] X. Zhou, Y. Jiang, and M. Bansal, "Data Factors for Better Compositional Generalization," *EMNLP 2023 - 2023 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 14549–14566, 2023, doi: 10.18653/v1/2023.emnlp-main.898.
[38] E. Ben-Zaken, S. Ravfogel, and Y. Goldberg, "BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2, pp. 1–9, 2022, doi: 10.18653/v1/2022.acl-short.1.
[39] D. Rustam *et al.*, "Development of Classification Method for Lecturer Area of Expertise Based on Scientific Publication Using BERT," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 3, pp. 894–905, 2024, doi: 10.18517/ijaseit.14.3.19893.

Vol. 6, No. 5, October 2025, Page. 5291-5304
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.4757

[40] G. Phillips *et al.*, "Setting nutrient boundaries to protect aquatic communities: The importance of comparing observed and predicted classifications using measures derived from a confusion matrix," *Sci. Total Environ.*, vol. 912, no. November 2023, 2024, doi: 10.1016/j.scitotenv.2023.168872.

[41] H. Shen, H. Jin, Á. A. Cabrera, A. Perer, H. Zhu, and J. I. Hong, "Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance," *Proc. ACM Human-Computer Interact.*, vol. 4, no. CSCW2, 2020, doi: 10.1145/3415224.

[42] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," 2020, doi: https://doi.org/10.48550/arXiv.2004.05150.

[43] Y. Hu, J. Ding, Z. Dou, and H. Chang, "Short-Text Classification Detector: A Bert-Based Mental Approach," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/8660828.

[44] O. Chew, H.-T. Lin, K.-W. Chang, and K.-H. Huang, "Understanding and Mitigating Spurious Correlations in Text Classification with Neighborhood Analysis," in *Findings of the Association for Computational Linguistics: EACL 2024*, Mar. 2024, pp. 1013–1025. doi: https://doi.org/10.48550/arXiv.2305.13654.

[45] M. G. Kim, M. Kim, J. H. Kim, and K. Kim, "Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter," *Int. J. Environ. Res. Public Health*, vol. 19, no. 9, 2022, doi: 10.3390/ijerph19095126.

[46] P.-J. Lin, R. Balasubramanian, F. Liu, and N. Kandpal, "Efficient Model Development through Fine-tuning Transfer," pp. 1–21, 2025, doi: https://doi.org/10.48550/arXiv.2503.20110.

[47] M. Bucos and G. Țucudean, "Text Data Augmentation Techniques for Fake News Detection in the Romanian Language," *Applied Sciences*, vol. 13, no. 13. 2023. doi: 10.3390/app13137389.