Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3356~3368
e-ISSN: 2548-6861

Headline-Based Indonesian Political Misinformation Classification: A Comparative Study of Naive Bayes, SVM, CNN, and IndoBERT Models

Ade Ariyo Yudanto [1]*, Naila Nabiha Qonita [2]*, Nur Aulia Maknunah [3]*, Inria Purwaningsih [4]*, Septian Rahardiantoro [5]*, Agus Mohamad Soleh [6]*
* Statistika dan Sains Data, IPB University
adeariyoyudanto@apps.ipb.ac.id [1], nailaqonita@apps.ipb.ac.id [2], nurauliamaknunah@apps.ipb.ac.id [3], inriapurwaningsih@apps.ipb.ac.id [4], septianrahardiantoro@apps.ipb.ac.id [5], agusms@apps.ipb.ac.id [6]

Article Info

Article history:

Received 2026-06-12
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Misinformation Classification,
Headline Classification,
IndoBERT,
Shortcut Learning,
Source Bias.

ABSTRACT

Political misinformation spreads rapidly through digital media, making early detection increasingly important. This study evaluates headline-based Indonesian political misinformation classification using Naive Bayes (NB), Support Vector Machine (SVM), Convolutional Neural Network (CNN), and IndoBERT. Headlines were selected because they represent the first information encountered by readers and enable rapid screening, although they may not fully represent the article content. The dataset consists of 5,132 Indonesian political news headlines collected from TurnBackHoax and DetikNews Politics. Leakage-aware preprocessing was applied to reduce explicit source-related cues before model training. Experimental results show that IndoBERT achieved the best performance with an accuracy of 90.25% and a Macro F1-score of 90.24%, outperforming the other evaluated models. Additional analyses, including McNemar statistical significance testing, computational cost comparison, and error analysis, showed that although IndoBERT achieved the highest predictive performance, its improvement over the other evaluated models was not statistically significant. Error analysis further revealed that several misclassification cases involved stylistic overlaps between misinformation and factual headlines, suggesting the possibility of residual source-style dependency and shortcut learning. Therefore, the proposed models should be interpreted as learning linguistic patterns associated with the constructed headline dataset rather than performing direct factual verification of news claims.

I. INTRODUCTION

Political misinformation has become a growing challenge in digital societies, particularly in countries with high social media penetration such as Indonesia. The rapid dissemination of misinformation, disinformation, and fake news through digital platforms can influence political attitudes, shape public opinion, and affect democratic processes. Social media platforms facilitate the rapid circulation of politically charged content, allowing misinformation to spread beyond traditional gatekeeping mechanisms. Previous studies have highlighted the role of social media in amplifying political engagement while simultaneously increasing exposure to misleading information [1]. Furthermore, misinformation has increasingly been recognized as a public communication problem with significant social and political consequences [2]. Within the Indonesian context, digital platforms have become important arenas for political discourse, making misinformation and disinformation persistent challenges to information integrity and public trust [3].

To address this problem, researchers have applied various Natural Language Processing (NLP) approaches for misinformation classification. Early studies commonly relied on traditional machine learning algorithms such as Naive Bayes (NB) and Support Vector Machine (SVM) using handcrafted textual features and TF-IDF representations. Subsequent developments introduced deep learning models such as Convolutional Neural Networks (CNN), which can

automatically capture local semantic patterns from text. More recently, transformer-based architectures have become the dominant paradigm due to their ability to learn contextual language representations. In the Indonesian language domain, transformer models such as IndoBERT have consistently demonstrated superior performance compared with conventional machine learning and earlier neural approaches. Previous studies have reported the effectiveness of BERT-based models for Indonesian fake news detection and political hoax classification, often outperforming CNN and other baseline models [4], [5], [6], [7].

Unlike full-text news articles, headlines represent the earliest textual information encountered by readers and are widely disseminated through news aggregators and social media platforms. Consequently, headline-based misinformation detection has practical value because it enables rapid content screening before users access the complete article. This is particularly relevant in social media environments, where many users form initial impressions and make sharing decisions based solely on headlines. Nevertheless, headline-based classification is inherently more challenging because headlines are often concise, ambiguous, sensational, or designed to attract attention without fully representing the article content. Therefore, the proposed model should be interpreted as learning linguistic patterns associated with misinformation headlines rather than performing direct fact verification of complete news articles.

Despite these advances, several methodological limitations remain. Existing studies frequently prioritize classification performance while paying limited attention to dataset construction and validity threats. Research has shown that models may learn stylistic characteristics of news sources rather than generalizable misinformation patterns, leading to source-style bias and reduced cross-domain robustness [8]. Shortcut learning occurs when models exploit superficial lexical or stylistic cues that are highly correlated with class labels instead of learning the underlying semantic characteristics of misinformation. Such shortcuts may arise from source-specific writing styles, platform-dependent language, or explicit source identifiers, resulting in overly optimistic performance without necessarily improving generalization. Dataset heterogeneity, platform-specific language use, and propagation-related biases further complicate misinformation classification tasks [9]. In addition, the concepts of misinformation, disinformation, and fake news are often used inconsistently, creating ambiguity in research objectives and interpretation of results [2]. Consequently, high classification performance should not be interpreted as evidence of factual reasoning or truth verification capability.

Based on these observations, several research gaps can be identified. First, relatively few studies focus specifically on Indonesian political misinformation using headline-only datasets. Second, source-style dependency and shortcut learning remain underexplored in Indonesian misinformation classification research. Third, comprehensive comparative evaluations involving NB, SVM, CNN, and IndoBERT within a consistent experimental framework are still limited. Finally, discussions regarding source asymmetry, platform bias, and methodological limitations are often underreported. Therefore, this study evaluates headline-based Indonesian political misinformation classification using NB, SVM, CNN, and IndoBERT on a heterogeneous dataset collected from TurnBackHoax and DetikNews Politics. Leakage-aware preprocessing was applied to reduce explicit source-related cues before model training. Nevertheless, because the dataset was constructed from two different news sources, residual source-specific linguistic patterns may still remain and are acknowledged as a limitation of this study. The contributions of this research include constructing a balanced political misinformation dataset, providing a comprehensive comparison of classical machine learning, deep learning, and transformer-based models, and presenting statistical significance analysis, computational cost evaluation, and error analysis to provide a more comprehensive assessment of headline-based political misinformation classification.

II. METHOD

A. Dataset & Exploratory Data Analysis

The dataset was constructed from two publicly accessible Indonesian online sources representing contrasting information categories. Misinformation samples were collected from TurnBackHoax, a fact-checking platform managed by MAFINDO that archives verified misinformation and hoax-related content. Valid news headlines were collected from the Politics section of DetikNews. Data collection and scraping were conducted between 10 and 15 May 2026. To reduce potential source leakage and improve comparability across sources, only news headlines were retained as model inputs, while article content was excluded from the final dataset.

The headline-only setting was intentionally adopted for three reasons. First, headlines represent the earliest and most frequently consumed textual information in online news environments and social media sharing practices. Second, using headlines improves comparability between misinformation and factual news sources because article structures, writing styles, and document lengths differ substantially across platforms. Third, preliminary experiments using full articles resulted in substantially higher performance, suggesting that article content may amplify source-specific patterns and increase the risk of source leakage. Therefore, headline-only classification was considered a more conservative and methodologically defensible setting for evaluating linguistic patterns associated with political misinformation.

The final dataset consists of 5,132 unique Indonesian political news headlines with six attributes: headline text (*judul*), class label (*label*), topic category (*topic*), word count

(*word_count*), tokenized text (*tokens_clean*), and preprocessed text (*text_clean*).

As shown in Figure 1, the class distribution is nearly balanced, comprising 2,592 valid headlines (50.51%) and 2,540 misinformation headlines (49.49%). To facilitate exploratory analysis, each headline was assigned a topic label using a rule-based tagging approach.

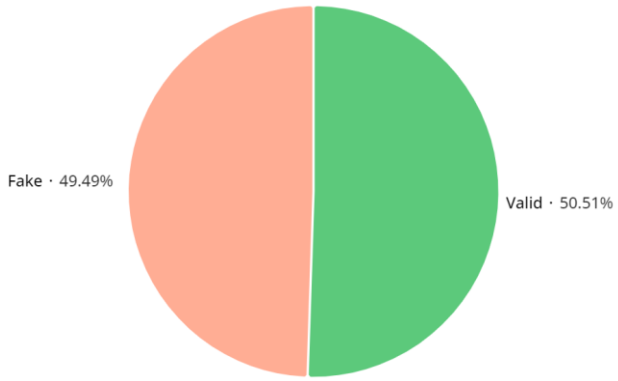

Figure 1. Label Distribution

Figure 2 presents the topic distribution, where government-related issues dominate the dataset (1,603 headlines), followed by miscellaneous political topics (1,492) and election-related topics (870). Other categories include international affairs (601), law and security (267), socio-economic issues (227), and education (72). These topic labels were used solely as descriptive metadata for dataset analysis and were not included as predictive features during model training.

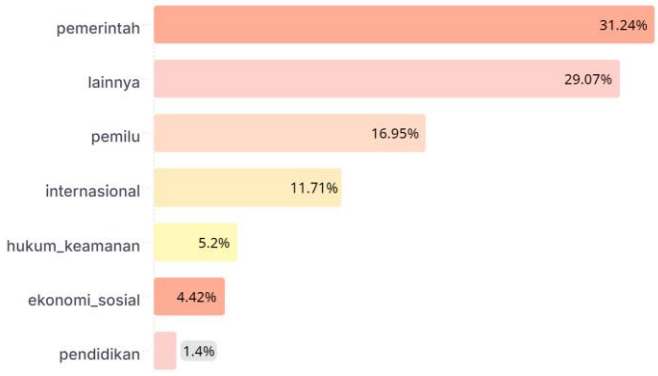

Figure 2. Topic Distribution

Figure 3 summarizes headline length statistics. The average headline contains 7.82 words with a standard deviation of 2.15 words. Headline lengths range from 3 to 29 words, while 75% of headlines contain nine words or fewer, indicating that most samples consist of short textual content suitable for headline-based classification.

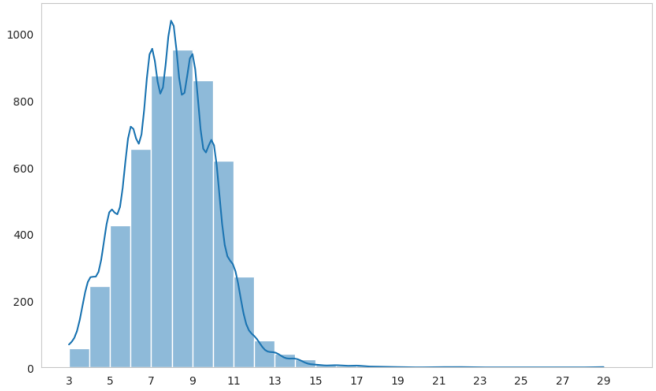

Figure 3. Word Count Histogram

A separate frequency analysis presented in Table I revealed that political figures such as Prabowo, Jokowi, Gibran, and Anies appeared frequently in the dataset, indicating that Indonesian political discourse is strongly associated with prominent political actors.

TABLE I
TOP FREQUENT TERMS

Word	Frequency
Prabowo	551
Jokowi	415
Gibran	299
Anies	292
Israel	290

B. *Data Preprocessing*

Data preprocessing was performed to improve text quality, reduce noise, and prepare the dataset for feature extraction and model training. Initially, only headline text was retained from both misinformation and factual news sources. The two datasets were then merged into a single corpus and assigned binary labels (fake and valid) for classification purposes.

Text normalization was applied by converting all headlines to lowercase, removing URLs, punctuation, special characters, and excessive whitespace. In addition, metadata markers such as [salah] and [misinformasi] were removed because they originated from the fact-checking platform and did not represent the actual headline content. These steps aimed to produce a standardized textual representation while preserving the semantic information contained in each headline.

To reduce source-related bias, several keywords that could directly reveal the origin of a headline were removed. These included *turnbackhoax*, *mafindo*, *salah*, *misinformasi*, and other platform-specific indicators commonly associated with fact-checking content. Additional terms related to social media engagement and source artifacts, such as viral, unggahan, akun, dibagikan, arsip, and video, were also removed. This procedure was designed to reduce explicit source indicators and encourage models to rely more on

linguistic patterns present in the headlines, although residual source-related characteristics may still remain.

Tokenization was performed using a whitespace-based approach (split()). Subsequently, stopword removal was applied using the Indonesian stopword list provided by Sastrawi [10], combined with a custom stopword collection specifically designed for Indonesian political news headlines. The custom stopwords included weak semantic terms, social-media-specific expressions, source identifiers, and other low-information tokens. Tokens shorter than three characters and purely numeric tokens were also removed to improve text quality.

Stemming was intentionally not applied in this study. Since the dataset consists of short political news headlines, preserving original word forms was considered important for maintaining contextual and entity-specific information, including names of political actors, institutions, and public events. Furthermore, the main objective of the study was comparative classification rather than linguistic normalization, making lightweight preprocessing more appropriate for evaluating the behaviour of classical machine learning and transformer-based models under realistic conditions.

After preprocessing, the remaining tokens were concatenated into a cleaned text representation (*text_clean*) used for model training. Word count statistics were then calculated for each headline, and topic labels were assigned using a rule-based keyword matching approach. Finally, duplicate headlines and samples containing fewer than three meaningful words were removed, resulting in a dataset of 5,132 unique headlines ready for subsequent feature extraction and classification experiments. Examples of the preprocessing results are presented in Table II.

Following preprocessing, the dataset was partitioned into three mutually exclusive subsets: training (80%), validation (10%), and test (10%), employing a two-stage stratified random sampling procedure with a fixed random seed of 42 to guarantee experimental reproducibility. Stratification was conducted based on a composite label-topic variable rather than class labels in isolation, thereby preserving the proportional distribution of both misinformation categories and topical domains across all resulting subsets. The training subset served as the basis for fitting the classification models, the validation subset was utilized for model selection and systematic hyperparameter optimization, and the test subset was held out exclusively for unbiased final performance assessment. This partitioning methodology was deliberately designed to mitigate sampling bias while ensuring a rigorous and consistent comparative evaluation across all candidate models.

TABLE II
BEFORE VS AFTER PREPROCESSING EXAMPLE HEADLINE

Before	After
[SALAH] Video "Tank Israel Diserang Pejuang Lebanon"	tank israel diserang pejuang lebanon
[SALAH] Korea Utara Bersiap Luncurkan 500 Rudal ke New York	korea utara bersiap luncurkan rudal new york
OPM Papua Paksa Warga Tanam Ganja di Halaman Rumah, Sasar Generasi Muda	opm papua paksa warga tanam ganja halaman rumah sasar generasi muda

C. Feature Representation

This study employs two feature representation strategies according to the characteristics of the classification models. Traditional machine learning models, namely Naive Bayes (NB) and Support Vector Machine (SVM), utilize Term Frequency-Inverse Document Frequency (TF-IDF) to convert textual data into numerical feature vectors. Meanwhile, deep learning models (CNN and IndoBERT) use dense vector representations that preserve semantic information, enabling contextual understanding of Indonesian political news headlines.

TF-IDF is a feature extraction method that transforms text into numerical vector representations by measuring the importance of a term within a document relative to the entire corpus [11], [12]. It consists of two components: Term Frequency (TF), which measures how often a term appears in a document, and Inverse Document Frequency (IDF), which penalizes terms that occur frequently across the corpus [11]. The final TF-IDF score is computed as [12]:

$$TF - IDF(t, d, D) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \times log \frac{N}{1 + df(t)}$$

where $f_{t,d}$is the frequency of term $t$in document d, $N$is the total number of documents, and $df(t)$is the number of documents containing term t.

In this study, TF-IDF features were extracted from the pre-processed headline text (*text_clean*) and used as the input representation for both the Naive Bayes and SVM classifiers. The TF-IDF vectorizer was configured using unigram and bigram features with a maximum vocabulary size of 6,111 terms to capture lexical patterns associated with political misinformation headlines. This representation provides an efficient and interpretable baseline for traditional machine learning models while maintaining computational efficiency [13].

In contrast to TF-IDF, deep learning models utilize dense vector representations capable of capturing richer semantic and contextual linguistic patterns. In this study, the CNN model employs an embedding layer that learns distributed word representations during training, allowing local contextual patterns within news headlines to be captured through convolution operations. Meanwhile, IndoBERT uses contextual embeddings generated by the pre-trained

indobenchmark/indobert-base-p1 model, enabling each word to be represented according to its surrounding context. Consequently, contextual representations are expected to provide better discrimination between factual and misinformation headlines than sparse lexical representations such as TF-IDF.

D. Model Architectures and Hyperparameter

Four classification models representing different learning paradigms were evaluated in this study, namely Naïve Bayes (NB), Support Vector Machine (SVM), Convolutional Neural Network (CNN), and IndoBERT. The TF-IDF configuration and the hyperparameter settings for the classical machine learning models are summarized in Tables III-V, while the CNN architecture and hyperparameter configuration are presented in Tables VI-VII. The architecture and fine-tuning configuration of IndoBERT are summarized in Tables VIII-IX.

TABLE III
HYPERPARAMETER CONFIGURATION FOR TF-IDF

Parameter	Configuration
Maximum Features	30000
N-gram Range	(1,2)
Minimum Document Frequency	2
Maximum Document Frequency	0.95
Sublinear TF Scaling	True

TABLE IV
HYPERPARAMETER CONFIGURATION FOR NAÏVE BAYES

Parameter	Configuration
Model	Multinomial Naive Bayes
Alpha	0.5
Fit Prior	True

TABLE V
HYPERPARAMETER CONFIGURATION FOR SVM

Parameter	Configuration
Model	LinearSVC
kernel	Linear
Regularization	1.0
Class Weight	balanced
Random State	42

TABLE VI
MODEL ARCHITECTURE FOR CNN

Layer	Specification	Configuration
Tokenization	Tokenizer	Keras Tokenizer
Embedding Layer	Input Dimension	15000
	Output Dimension	128
	Input Length	40
Convolution Layer	Activation	ReLU
	Filters	128
	Kernel Size	3
Pooling Layer	Global Max Pooling	True
Regularization	Dropout	0.5
Loos Function	Loss	Binary Crossentropy
Dense Layer	Hidden Units	128 Units, ReLU
Output Layer	Dense Units	1 Unit, Sigmoid

TABLE VII
HYPERPARAMETER CONFIGURATION FOR CNN

Parameter	Configuration
Max Words	15000
Max Length	40
Batch Size	32
Epoch	15
Learning Rate	0.001
Dropout Rate	0.5
Max Sequence Length	40
Optimizer	Adam
Validation Set Ratio	0.1

TABLE VIII
MODEL ARCHITECTURE FOR INDOBERT

Layer	Specification	Configuration
Input Layer	Max Sequence Length	64
Tokenization	Tokenizer	AutoTokenizer
Pre-trained Model	Model	*indobenchmark/ indobert-base-p1*
Transformer Encoder	Hidden Size	768
	Number of Layers	12
	Attention Heads	12
Classification Head	Number of Labels	2
Output Layer	Output	Logits / Softmax for prediction

TABLE IX
HYPERPARAMETER CONFIGURATION FOR INDOBERT

Parameter	Configuration
Max Sequence Length	64
Batch Size	16
Epoch	3
Learning Rate	2e-5
Dropout Rate	0.1 (default model configuration)
Optimizer	*AdamW*
Weight Decay	0.01

The two classical machine learning models, Naïve Bayes and SVM, shared the same TF-IDF feature representation and preprocessing pipeline to ensure that performance differences were attributable to the classifiers rather than the input features. In contrast, CNN and IndoBERT directly learned semantic representations from the preprocessed headlines through trainable embedding mechanisms and contextual language modeling, respectively.

The CNN architecture consisted of an embedding layer followed by a one-dimensional convolutional layer, global max pooling, dropout regularization, and a fully connected output layer. Meanwhile, IndoBERT was fine-tuned from the pre-trained *indobenchmark/indobert-base-p1* model using the AdamW optimizer and a low learning rate to preserve the knowledge acquired during pre-training. Early stopping based on validation performance was employed for both deep learning models to reduce overfitting and retain the best-performing model.

To ensure a fair comparison, all models were trained and evaluated using the same training, validation, and test partitions as well as identical evaluation metrics. The selected hyperparameters were determined based on preliminary experiments and commonly adopted settings reported in previous studies.

E. Naïve Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem that estimates the posterior probability of each class from the observed features. The method assumes conditional independence among features, allowing efficient computation even in high-dimensional text data. Despite its simplicity, Naive Bayes remains one of the most widely used approaches for text classification due to its computational efficiency, robustness, and competitive performance across various natural language processing tasks [14].

In this study, TF-IDF feature vectors were used as input to a Multinomial Naive Bayes classifier. The Multinomial variant was selected because it is well suited for text representation based on term frequencies and has indicated effective performance in handling high-dimensional textual data [14]. Model development and prediction were implemented using the *MultinomialNB* algorithm available in the scikit-learn library [15]. During classification, the model calculates the posterior probability of each class and assigns the class with the highest probability as the prediction result.

F. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm widely used for text classification tasks due to its ability to effectively handle high-dimensional data and perform well on sparse feature representations such as TF-IDF [16], [17]. Studies on text categorization have indicatted that SVM achieves strong classification performance when documents are represented as high-dimensional feature vectors [16]. More recent reviews have also identified SVM as one of the most widely used machine learning algorithms for text classification and misinformation classification tasks [17], [18].

The fundamental principle of SVM is to capture an optimal hyperplane that separates two classes with the maximum possible margin [19]. The margin is defined as the distance between the hyperplane and the nearest data points from each class. These nearest data points, known as support vectors, play a critical role in determining the position of the model's decision boundary [19].

In this study, a Linear SVM model was employed because it is well suited for text data represented as high-dimensional and sparse TF-IDF vectors [16], [17]. Furthermore, recent studies on misinformation classification and political headline classification have reported that the combination of TF-IDF and SVM remains a competitive baseline despite the growing adoption of deep learning and transformer-based approaches [18], [20].

G. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is one of the deep learning architectures widely used in text classification tasks because of its ability to effectively extract important local patterns from word sequences [21]. In this study, CNN was utilized as a deep learning baseline model for headline classification, where Indonesian political news headlines were categorized into fake-labelled and valid-labelled classes.

Before the model training process, the news headlines underwent several preprocessing stages, including case folding, cleaning, tokenization, stopword removal, and sequence padding. After preprocessing, the headlines were transformed into numerical representations using an embedding layer to enable the model to learn semantic relationships between words [22].

CNN works by applying convolution layers to capture important patterns within news headlines, such as sensational phrasing, clickbait expressions, and specific word combinations that frequently occur in headlines labeled as misinformation within the constructed dataset [23], [24]. Subsequently, the extracted features were selected using Global Max Pooling before being forwarded to Dense layers for binary classification.

To reduce overfitting during training, dropout regularization was implemented [25]. In addition, the Adam optimizer was utilized because of its adaptive and stable learning capability in deep learning optimization tasks [26].

H. IndoBERT

IndoBERT is a Transformer-based language model specifically developed for Indonesian Natural Language Processing (NLP) tasks [27]. The model utilizes a self-attention mechanism to understand bidirectional contextual relationships between words, allowing it to capture sentence meaning more effectively than traditional sequential models [28].

In this study, the *indobenchmark/indobert-base-p1* model was employed as the primary model for headline classification by assigning Indonesian political news headlines into misinformation-labelled and valid-labelled categories. Compared to CNN, IndoBERT provides better contextual understanding, which may provide advantages when processing ambiguous, provocative, and semantically complex headlines [29].

The preprocessing stage in IndoBERT employed the WordPiece tokenizer to split words into sub word tokens, enabling the model to handle vocabulary variations and unseen words more efficiently [28]. After tokenization, the headlines were transformed into contextual embeddings before undergoing the fine-tuning process using the pre-processed political headline dataset.

To improve training stability, this study utilized the AdamW optimizer and a relatively low learning rate to preserve the knowledge obtained during the pre-training stage [30]. In addition, the maximum sequence length was limited to 128 tokens to improve computational efficiency while maintaining important contextual information from the headlines.

I. Evaluation Metrics

The purpose of model evaluation is to measure the effectiveness and overall quality of a model. This includes accuracy, representational capability, relevance of similarity, appropriateness for the task, and other performance-related criteria [31]. In this study, model performance was evaluated using a confusion matrix, which is composed of four fundamental elements: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These components form the basis for the following evaluation metrics:

1) Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

3) Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

4) F1-Score

$$\text{F1-Score} = 2 \times \frac{Recall \times \text{Precision}}{Recall + Precision}$$

J. Statistical Significance Analysis

To determine whether the performance differences among the classification models were statistically significant, McNemar's test was employed. McNemar's test is a non-parametric statistical test designed for paired nominal data and is commonly used to compare the prediction outcomes of two classification models evaluated on the same test dataset [32]. Unlike performance metrics such as Accuracy or Macro F1-score, McNemar's test assesses whether the observed differences in prediction errors between two models are statistically significant rather than arising from random variation.

In this study, pairwise comparisons were conducted among the four classification models (Naive Bayes, SVM, CNN, and IndoBERT) using the predictions obtained from the same test set. The null hypothesis states that the two models have the same error rate, whereas the alternative hypothesis assumes that their error rates differ significantly. Statistical significance was determined using a significance level of $\alpha = 0.05$, where a p-value less than 0.05 indicates a statistically significant difference between the compared models.

III. RESULT AND DISCUSSION

A. Model Performance Analysis

TABLE X
CLASS-WISE PERFOMANCE

Model	Metric	Class	
		Valid	Fake
TF-IDF + Naïve Bayes	Precision	89.84%	85.39%
	Recall	85.00%	90.12%
	F1-Score	87.35%	87.69%
TF-IDF + SVM	Precision	89.75%	84.76%
	Recall	84.23%	90.12%
	F1-Score	86.90%	87.36%
CNN	Precision	88.84%	85.88%
	Recall	85.77%	88.93%
	F1-Score	87.28%	87.38%
IndoBERT	Precision	93.75%	87.18%
	Recall	86.54%	94.07%
	F1-Score	90.00%	90.49%

Table X summarizes the class-wise precision, recall, and F1-score obtained by each evaluated model for the valid and misinformation classes. Overall, the experimental results indicate that all models achieved strong classification performance, with accuracy values exceeding 87%, suggesting that political news headlines contain sufficiently discriminative linguistic characteristics under the constructed experimental setting. Among the four approaches, IndoBERT achieved the highest accuracy of 90.25%, outperforming all other models by a considerable margin. The remaining three models produced relatively similar results, with Naïve Bayes achieving an accuracy of 87.52%, followed by CNN at 87.32% and SVM at 87.13%. The relatively small performance differences among these three approaches suggest that headline information alone provides useful discriminatory cues for separating the two headline categories in the constructed dataset.

The superior performance of IndoBERT suggests that contextual language representations may provide advantages for headline classification within the political news dataset constructed in this study. Unlike TF-IDF-based approaches that primarily rely on word occurrence patterns, IndoBERT leverages bidirectional contextual representations that enable the model to capture semantic relationships among words. This capability may be advantageous when processing political headlines in which some misinformation-labelled headlines contain ambiguous phrasing, implicit claims, or persuasive language that cannot be represented solely through isolated keywords. As a result, IndoBERT achieved the

highest classification performance in this experiment. However, the present study does not establish whether the performance gains originate from deeper semantic representations, residual source-specific cues, or a combination of both.

An interesting finding is the competitive performance of the TF-IDF-based models. Despite their relatively simple architecture, both Naïve Bayes and SVM achieved performance levels that were only slightly lower than CNN. In particular, Naive Bayes obtained the best result among the non-transformer approaches. This outcome suggests that lexical characteristics observed in the collected headlines provide useful discriminatory signals for the classification task defined in the constructed dataset. Since headlines are generally short and concise, discriminative keywords and phrase patterns captured by TF-IDF may provide useful information for separating the two headline categories.

Another notable observation is that CNN did not achieve a substantial advantage over the traditional machine learning models. Although CNN can learn local semantic patterns through distributed word representations, its performance remained very close to that of both Naïve Bayes and SVM. This result may indicate that the relatively short length of news headlines limits the benefits typically associated with deep neural architectures. When the classification task is largely driven by distinctive lexical signals, simpler models can remain highly competitive despite their lower complexity.

Overall, the findings indicate that contextual language modeling achieved the best classification performance under the headline-based experimental setting adopted in this study, as evidenced by the clear performance advantage of IndoBERT. Nevertheless, the strong performance achieved by the TF-IDF-based models suggests that traditional machine learning approaches remain viable alternatives, particularly in scenarios where computational efficiency, model simplicity, and ease of deployment are important considerations.

B. Confusion Matrix

Figures 4-7 present the confusion matrices obtained for the Naïve Bayes, SVM, CNN, and IndoBERT models, respectively, illustrating the distribution of correct and incorrect predictions on the test set. These confusion matrices provide a more detailed view of each model's classification behaviour beyond the overall evaluation metrics by highlighting the patterns of false positives (FP) and false negatives (FN).

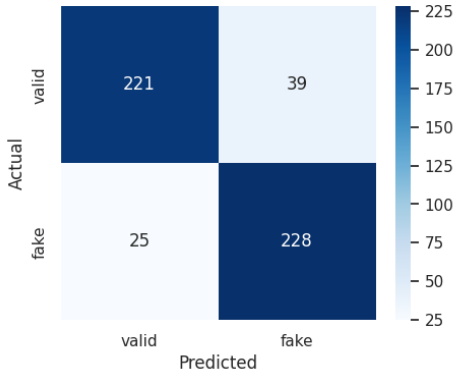

Figure 4. Confusion Matrix Naive Bayes + TF-IDF

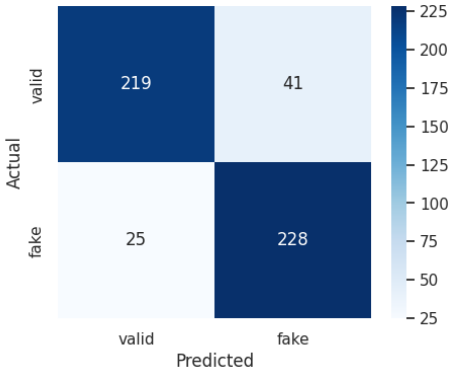

Figure 5. Confusion Matrix SVM + TF-IDF

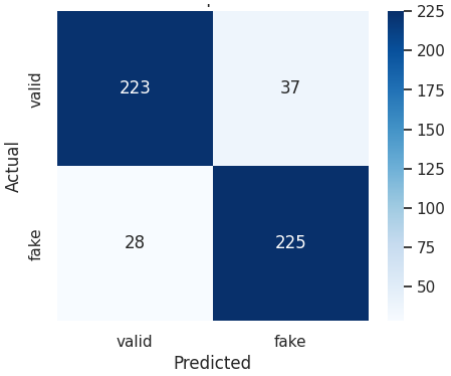

Figure 6. Confusion Matrix CNN

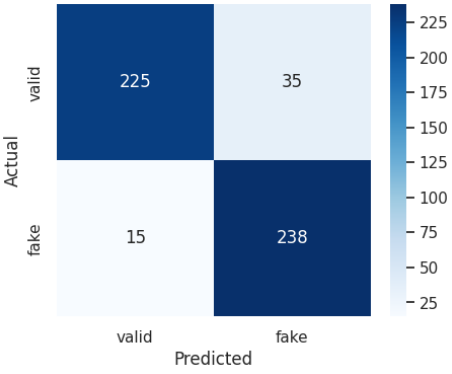

Figure 7. Confusion Matrix IndoBERT

The *confusion matrix* results on the *test* data reveal more specific misclassification patterns for each model. Naïve Bayes and SVM exhibited nearly identical error patterns, where both models produced the same FN of 25 samples, while the FP of Naive Bayes at 39 was slightly lower than that of SVM at 41. This similarity is consistent with the fact that both models rely on TF-IDF representations, resulting in comparable limitations in distinguishing headlines from the two labeled categories when they share similar lexical patterns.

CNN displayed a slightly different pattern, where FP decreased to 37 but FN increased to 28. This trade-off suggests that CNN tends to be more conservative in predicting the fake class, but as a consequence misses more actual fake samples. IndoBERT recorded the most balanced error distribution, with the lowest FN of 15 samples and an FP of 35 samples. The lower FN observed for IndoBERT suggests that the model was comparatively more effective at classifying headlines assigned to the misinformation-labelled category within the constructed dataset.

Overall, all models tend to produce higher FP than FN, suggesting that a portion of valid headlines share lexical characteristics that overlap with headlines labelled as misinformation in the dataset. This observation is consistent with the possibility of source-style bias discussed in this study, although additional experiments would be required to verify this explanation empirically.

C. Shortcut Learning and Source-Style Dependency Analysis

Shortcut learning refers to a phenomenon in which machine learning models rely on superficial patterns or spurious correlations that happen to be associated with class labels rather than learning the intended underlying concepts. Such shortcuts often produce strong performance on in-distribution data but lead to reduced robustness and limited generalization under distribution shifts or real-world conditions. Previous studies have shown that modern machine learning models, including deep neural networks and language models, frequently exploit dataset artifacts and source-specific cues because these signals provide easier decision rules than learning the actual semantic task of interest [33], [34], [35].

In the present study, misinformation and factual headlines originated from two different online platforms, namely TurnBackHoax and DetikNews Politics. Although leakage-aware preprocessing was implemented by removing explicit source indicators and fact-checking markers, residual differences in writing style, lexical preferences, headline structure, and platform conventions may still remain. Consequently, the evaluated models may partially exploit these source-specific characteristics as predictive shortcuts instead of relying solely on linguistic patterns that generalize beyond the constructed dataset. Similar concerns have been reported in misinformation classification research, where models can inadvertently learn source-dependent artifacts and exhibit reduced robustness when evaluated under different domains or distributions [8], [9].

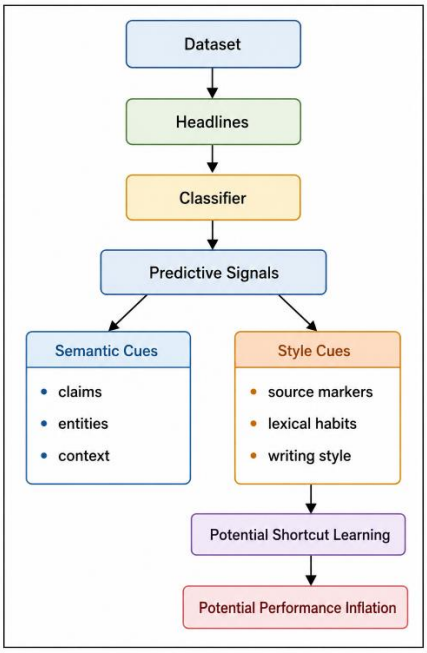

Figure 8. Semantic and Source-Style Signals in Headline Classification

Figure 8 illustrates the conceptual distinction between semantic cues and style cues that may be exploited during headline classification. Semantic cues refer to information related to claims, entities, and contextual expressions contained in headlines. In contrast, style cues comprise source-specific characteristics such as lexical preferences, writing conventions, and platform-dependent expressions. Under source-asymmetric settings, models may partially rely on these stylistic signals because they provide simpler predictive rules, potentially leading to shortcut learning and inflated in-distribution performance.

In this study, shortcut learning was operationally interpreted as the possibility that classification performance was partially driven by source-related artifacts rather than purely semantic understanding. Four empirical indicators were considered: (1) consistently high performance across all evaluated models, including relatively simple classifiers; (2) competitive performance of TF-IDF-based approaches despite their limited contextual representation capability; (3) substantially higher performance observed during preliminary experiments using complete news articles instead of headlines alone; and (4) highly similar error patterns across different model architectures. A summary of these indicators and their possible interpretations is presented in Table XI. Although these observations do not constitute direct proof of shortcut learning, they provide convergent evidence that source-related cues may contribute to model predictions.

TABLE XI
INDICATORS OF POTENTIAL SHORTCUT LEARNING AND SOURCE-STYLE DEPENDENCY

Observation	Empirical Evidence	Possible Interpretation
High performance across all models	NB, SVM, CNN, and IndoBERT achieved accuracy above 87%	The dataset may contain easily exploitable predictive cues
Competitive performance of SVM	Small performance gap between SVM and IndoBERT	Surface lexical patterns remain highly informative
Higher performance using complete articles	Preliminary experiments outperformed the headline-only setting	Additional source-related signals may be present in longer texts
Similar confusion matrices	Comparable error distributions across architectures	Different models may rely on overlapping predictive cues

Collectively, these findings suggest the possibility of source-style dependency and shortcut learning within the constructed dataset. Therefore, the reported performance should be interpreted cautiously because part of the predictive capability may arise from source-related patterns in addition to genuine linguistic characteristics associated with misinformation-labelled and valid-labelled headlines. Consequently, the findings of this study are better interpreted as evidence of effective headline classification under the constructed experimental setting rather than as direct evidence of factual reasoning or real-world misinformation verification capability. Future studies should employ external datasets and cross-source evaluation protocols to empirically examine the extent of shortcut learning and assess model robustness under distribution shifts.

The present study does not empirically quantify shortcut learning through cross-source evaluation, leave-one-source-out validation, or external dataset evaluation. Consequently, source-style dependency should be interpreted as a plausible explanation supported by convergent experimental observations rather than as a directly verified causal mechanism.

D. Error Analysis

To better understand the limitations of the proposed classification models, an error analysis was conducted on the predictions produced by the best-performing model (IndoBERT). Rather than relying solely on aggregate performance metrics, this analysis examined representative misclassified examples to identify common characteristics and potential sources of classification errors. Representative examples are presented in Table XII, while the main error categories are summarized in the accompanying discussion.

TABLE XII
REPRESENTATIVE ERROR ANALYSIS OF MISCLASSIFIED HEADLINES

Error Category	Representative Headline	Possible Explanation
Formal misinformation resembling factual news	*Kesuksesan Food Estate Prabowo di Jawa Tengah*	Resembled factual news after preprocessing
Sensational factual headline	*Heboh Barang Pemberian China Dibuang Rombongan Trump!*	Attention-grabbing wording resembled stylistic patterns commonly found in misinformation headlines.
Shared political entities	*Bunda Iffet Meninggal Dunia, Ganjar Pranowo-Armand Maulana Melayat ke Potlot*	Both classes frequently mentioned the same public figures, reducing lexical discriminability
Very short headline	*Algoritma Kebijakan Prabowo*	Limited contextual information made semantic interpretation more difficult

Error analysis revealed that most classification errors occurred when the linguistic characteristics of misinformation and factual headlines overlapped. Several misinformation headlines were written in a neutral journalistic style after preprocessing removed explicit misinformation indicators, making them difficult to distinguish from factual news. Conversely, a few factual headlines employed sensational wording that resembled stylistic patterns commonly associated with misinformation headlines. In addition, many misclassified headlines referred to the same political figures and international issues, indicating that topical overlap between the two classes further reduced the discriminative power of lexical features. Only a small proportion of errors involved very short headlines, suggesting that headline length was not the primary source of misclassification.

E. Practical Implications and Model Trade-Off

Although IndoBERT achieved the highest overall performance with a Macro F1-score of 90.24%, the improvement over SVM was relatively modest, amounting to only 3.11 percentage points. While the gains compared to CNN and Naïve Bayes were more pronounced, they were also accompanied by significantly higher computational demands. These results suggest that model performance should not be evaluated solely based on predictive accuracy, but also in relation to model complexity and the practical costs associated with training and deployment.

TABLE XII
MODEL PERFORMANCE AND COMPUTATIONAL TRADE-OFF

Model	Accuracy (%)	Training Time (s)	Inference Time (s)	Complexity
TF-IDF + NB	87.52	0.2195	0.0206	6,111 TF-IDF features
TF-IDF + SVM	87.13	0.1979	0.0121	6,111 TF-IDF features
CNN	87.32	14.5055	0.06031	1,985,921 parameters
Indo BERT	90.25	104.0068	0.6995	124,442,882 parameters

Table XII demonstrates a clear trade-off between predictive performance and computational requirements. Although IndoBERT achieved the highest accuracy (90.25%), its improvement over the other models was relatively modest, ranging from approximately 2.7 to 3.1 percentage points. In contrast, IndoBERT required substantially longer training and inference times and exhibited the highest model complexity, comprising more than 124 million parameters. Conversely, TF-IDF-based approaches, particularly SVM, achieved competitive performance while requiring only a fraction of the computational cost. These findings suggest that model selection for political misinformation classification should consider not only predictive performance but also operational constraints, deployment feasibility, and computational efficiency. From a practical standpoint, SVM remains an attractive alternative because an accuracy difference of approximately three percentage points may not justify an increase of several hundred times in training cost and substantially slower inference in resource-constrained environments.

F. Statistical Significance Analysis

To examine whether the observed performance differences among the classification models were statistically significant, pairwise comparisons were conducted using McNemar's test. This test evaluates whether two classifiers exhibit significantly different prediction errors when applied to the same test instances. The results of the pairwise comparisons are presented in Table XII.

TABLE XIII
PAIRWISE MCNEMAR TEST RESULTS

Model Comparison	McNemar Statistic	p-value	Significant ($\alpha = 0.05$)
Naive Bayes vs. SVM	0.0250	0.8744	No
Naive Bayes vs. CNN	0.0000	1.0000	No
Naive Bayes vs. IndoBERT	2.7258	0.0987	No
SVM vs. CNN	0.0000	1.0000	No
SVM vs. IndoBERT	3.6290	0.0568	No
CNN vs. IndoBERT	3.6981	0.0545	No

Table XIII presents the pairwise McNemar test results for all classification models. McNemar's test was conducted to determine whether the observed differences in prediction performance between two models were statistically significant when evaluated on the same test dataset. All pairwise comparisons produced *p*-values greater than the significance threshold of 0.05, indicating that no statistically significant differences were observed among the four classification models.

Although IndoBERT achieved the highest Accuracy and Macro F1-score, the improvement over Naive Bayes, SVM, and CNN was not statistically significant at the 5% significance level. This finding suggests that the performance gains achieved by the transformer-based model were relatively modest on the evaluated dataset. Therefore, the selection of an appropriate classification model should also consider computational efficiency and deployment requirements in addition to predictive performance. These findings are consistent with the computational cost analysis, where simpler models such as Naive Bayes and SVM required substantially lower training and inference time than IndoBERT.

The absence of statistically significant differences indicates that the observed performance improvements should be interpreted cautiously, particularly considering the relatively limited size and domain-specific characteristics of the headline dataset.

G. Research Limitations

Despite the strong classification performance, several limitations should be acknowledged. News headlines are inherently short and often employ ambiguous, provocative, or clickbait-oriented expressions that may not accurately represent the complete article content. Consequently, the linguistic patterns learned by the models should not be interpreted as representations of factual correctness and may only capture a limited portion of the contextual information surrounding the underlying news events. Furthermore, because only headlines were analyzed, the present study should be interpreted as headline classification rather than comprehensive misinformation verification.

IV. CONCLUSION

This study compared the performance of Naïve Bayes, SVM, CNN, and IndoBERT for headline-based Indonesian political misinformation classification using a dataset collected from TurnBackHoax and DetikNews Politics. The results show that all evaluated models achieved strong performance, suggesting that the collected political news headlines contain sufficiently discriminative linguistic characteristics to support classification under the constructed experimental setting. Among the evaluated approaches, IndoBERT achieved the best overall performance with an accuracy of 90.25% and an F1-score of 90.24%. However, McNemar's test indicated that these improvements were not statistically significant compared with the other evaluated models. Therefore, the present study does not establish whether the observed performance gains originate from deeper semantic representations, residual source-specific cues, or a combination of both.

The findings also indicate that simpler approaches such as Naïve Bayes, SVM, and CNN remain highly competitive, suggesting that lexical and stylistic cues contribute meaningfully to headline classification. Furthermore, the error analysis revealed that most misclassification cases occurred when misinformation headlines closely resembled factual news reporting or when factual headlines adopted sensational writing styles, highlighting the challenges posed by semantic and stylistic overlap between the two classes. These observations, together with the competitive performance of lexical-based models, suggest the possibility of source-style dependency and shortcut learning within the constructed dataset despite the application of leakage-aware preprocessing.

Therefore, the reported performance should be interpreted as the effectiveness of headline classification within the constructed dataset rather than as a direct measure of real-world fact-checking capability. Future research should incorporate more diverse news sources, external benchmark datasets, and cross-source evaluation protocols to better assess model generalization and empirically examine the extent of source-related bias and shortcut learning in headline-based political misinformation classification before such models are deployed in practical misinformation detection systems.

ACKNOWLEDGEMENTS

Authors would like to thank Statistics and Data Science, IPB University, Bogor, Indonesia, for the institutional and research support provided throughout this study. The support contributed significantly to the completion of the research and manuscript preparation.

REFERENCES

[1] S. Valenzuela, D. Halpern, J. E. Katz, and J. P. Miranda, "The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation," *Digital Journalism*, vol. 7, no. 6, pp. 802–823, 2019.

[2] E. Broda and J. Strömbäck, "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review," *Ann. Int. Commun. Assoc.*, vol. 48, no. 2, pp. 139–166, 2024.

[3] D. Subekti, M. Yusuf, M. Saadah, and M. Wahid, "Social media and disinformation for candidates: the evidence in the 2024 Indonesian presidential election," *Front. Polit. Sci.*, vol. 7, p. 1625535, 2025.

[4] 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.

[5] S. F. N. Azizah, H. D. Cahyono, S. W. Sihwi, and W. Widiarto, "Performance analysis of transformer based models (BERT, ALBERT, and RoBERTa) in fake news detection," in *2023 6th International Conference on Information and Communications Technology (ICOIACT)*, IEEE, 2023, pp. 425–430.

[6] A. Simanjuntak *et al.*, "Research and analysis of indobert hyperparameter tuning in fake news detection," *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, vol. 13, no. 1, pp. 60–67, 2024.

[7] B. Prastyapradipta, K. Naoko, R. A. B. Himawan, M. A. Ibrahim, and G. A. Tarigan, "Analyzing Indonesian political hoax detection system on social media using deep learning and natural language processing," *Procedia Comput. Sci.*, vol. 269, pp. 1742–1751, 2025.

[8] C. Blackledge and A. Atapour-Abarghouei, "Transforming fake news: Robust generalisable news classification using transformers," in *2021 IEEE international conference on big data (big data)*, IEEE, 2021, pp. 3960–3968.

[9] T. Li, Y. Sun, S. Hsu, Y. Li, and R. C.-W. Wong, "Fake News Detection with Heterogeneous Transformer," May 2022, [Online]. Available: http://arxiv.org/abs/2205.03100

[10] F. Z. Tala, "Sastrawi: Indonesian Stemmer and Stopword Remover," GitHub.

[11] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. in International student edition. McGraw-Hill, 1983. [Online]. Available: https://books.google.co.id/books?id=7f5TAAAAMAAJ

[12] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing, 2008.

[13] I. Nyoman, P. Trisna, I. Made, W. J. Putra, and W. O. Vihikan, "Classifying Indonesian Hoax News Titles with SVM, XGBoost, and BiLSTM," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems*, vol. 19, no. 4, pp. 361–370, 2025, doi: 10.22146/ijccs.106608.

[14] A. McCallum, "A comparison of event models for naive bayes text classification," 1998.

[15] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[16] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, Springer, 1998, pp. 137–142.

[17] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, "Twenty years of machine-learning-based text classification: A systematic review," *Algorithms*, vol. 16, no. 5, p. 236, 2023.

[18] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," *Multimed. Tools Appl.*, vol. 83, no. 17, pp. 51009–51067, 2024, doi: 10.1007/s11042-023-17470-8.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[20] M. A. B. Al-Tarawneh, O. Al-irr, K. S. Al-Maaitah, H. Kanj, and W. H. F. Aly, "Enhancing fake news detection with word embedding: A machine learning and deep learning approach," *Computers*, vol. 13, no. 9, p. 239, 2024.

[21] S. M. Al-Selwi *et al.*, "RNN-LSTM: From applications to modeling techniques and beyond—Systematic review," *Journal of*

King Saud University - Computer and Information Sciences, vol. 36, no. 5, p. 102068, 2024, doi: https://doi.org/10.1016/j.jksuci.2024.102068.

[22] M. Waqas and U. W. Humphries, "A critical review of RNN and LSTM variants in hydrological time series predictions," *MethodsX*, vol. 13, p. 102946, 2024, doi: https://doi.org/10.1016/j.mex.2024.102946.

[23] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet – A deep convolutional neural network for fake news detection," *Cogn. Syst. Res.*, vol. 61, pp. 32–44, 2020, doi: https://doi.org/10.1016/j.cogsys.2019.12.005.

[24] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1746–1751.

[25] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh tahapan preprocessing terhadap model Indobert dan Indobertweet untuk mendeteksi emosi pada komentar akun berita Instagram," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, pp. 887–894, 2024.

[26] K. K. T. G, A. R, S. G. Koolagudi, T. Rao, and A. Kodipalli, "Stratification of Depressed and Non-Depressed Texts from Social Media using LSTM and its Variants," *Procedia Comput. Sci.*, vol. 235, pp. 1353–1363, 2024, doi: https://doi.org/10.1016/j.procs.2024.04.127.

[27] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857.

[28] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[29] W. A. Hidayat and V. R. S. Nastiti, "Perbandingan kinerja pre-trained IndoBERT-base dan IndoBERT-lite pada klasifikasi sentimen ulasan TikTok Tokopedia Seller Center dengan model IndoBERT," *JSiI (Jurnal Sistem Informasi)*, vol. 11, no. 2, pp. 13–20, 2024.

[30] S.-V. Oprea and A. Bâra, "Extracting Emotions from Customer Reviews Using Text Mining, Large Language Models and Fine-Tuning Strategies," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 3, p. 221, 2025.

[31] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[32] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[33] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[34] V. Dogra, S. Verma, M. Woźniak, J. Shafi, and M. F. Ijaz, "Shortcut learning explanations for deep natural language processing: A survey on dataset biases," *IEEE Access*, vol. 12, pp. 26183–26195, 2024.

[35] K. Hermann, H. Mobahi, T. Fel, and M. Mozer, "On the foundations of shortcut learning," in *International Conference on Learning Representations*, 2024, pp. 43832–43868.