International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

Comparative Study of TF-IDF-SVM and IndoBERT for Imbalanced Indonesian Hoax Detection

Yono Cahyono[*], Nardiono, Hendri Ardiansyah, Cintia Septiani

Faculty of Computer Science, Informatics Engineering Study Program, Universitas Pamulang, South Tangerang, Indonesia
Email: [1,*]dosen00843@unpam.ac.id, [2]dosen00834@unpam.ac.id, [3]dosen00830@unpam.ac.id, [4]cintiaseptiani@gmail.com
Correspondence Author Email: dosen00843@unpam.ac.id
Received: 12/05/2026, Revised: 01/06/2026, Accepted: 18/06/2026, Published: 30/06/2026

Abstract–Indonesian hoax detection remains challenging because classification performance is influenced by text representation, class imbalance, and experimental design. This study comparatively evaluates a conventional Term Frequency–Inverse Document Frequency with Support Vector Machine (TF-IDF–SVM) approach and the transformer-based IndoBERT model under the same experimental protocol on a naturally imbalanced Indonesian hoax dataset. News titles and narratives were combined as textual input, duplicate records were removed, and the resulting dataset was stratified into training, validation, and held-out test sets. Model configurations were selected exclusively using validation Macro F1-score to prevent test-set information from influencing model selection. The final SVM employed unigram–bigram TF-IDF features with balanced LinearSVC, whereas the selected IndoBERT configuration used unweighted cross-entropy loss with a maximum sequence length of 128 tokens. On the test set, TF-IDF–SVM achieved an accuracy of 0.8299, Macro Precision of 0.7123, Macro Recall of 0.7065, and Macro F1-score of 0.7093, while IndoBERT achieved 0.8047, 0.6665, 0.6573, and 0.6616, respectively. The stronger numerical performance of TF-IDF–SVM suggests that discriminative lexical patterns remained highly informative in this dataset, while the contextual advantages of IndoBERT may not have been fully realized under the substantial class imbalance and the investigated training configuration. However, the exact McNemar test produced a p-value of 0.1371, indicating that the difference in paired classification correctness was not statistically significant at the 0.05 level. These findings demonstrate that TF-IDF–SVM remains a competitive and computationally simpler baseline for Indonesian hoax detection under imbalanced data conditions, while the benefit of more complex contextual models remains dependent on dataset characteristics and experimental settings.

Keywords: Indonesian Hoax Detection; TF-IDF; Support Vector Machine; IndoBERT; Fake News Classification; Class Imbalance

1. INTRODUCTION

The rapid development of digital media has fundamentally changed how information is produced, distributed, and consumed. Online news platforms, social media, and messaging applications enable information to reach large audiences within a short period of time. However, the same accessibility also facilitates the rapid dissemination of fabricated, misleading, or unverifiable information. Fake news can influence public perception, weaken trust in information sources, and create social and political consequences, making automatic detection an increasingly important research problem [1], [2]. In Indonesia, the problem is particularly relevant because online misinformation occurs across various domains, including health, social issues, and politics. Recent research on Indonesian political misinformation has emphasized that hoaxes can threaten public trust, social stability, and evidence-based decision making [3], while fake-news content related to the Indonesian presidential election has also become an important subject of computational linguistic analysis [4].

Natural Language Processing (NLP) and machine learning provide scalable alternatives to manual fact-checking for identifying potentially misleading content. Traditional text-classification approaches generally convert textual documents into numerical representations such as Term Frequency-Inverse Document Frequency (TF-IDF), followed by classifiers including Support Vector Machine (SVM), Naïve Bayes, Random Forest, and Logistic Regression. These methods remain relevant because they are computationally efficient and can perform competitively on text-classification tasks. Zhou and Cai [5]noted that machine-learning methods such as SVM can provide fast prediction with relatively low resource consumption, although sparse representations such as TF-IDF have limited capability to explicitly represent deeper semantic relationships. Datta et al. [2] similarly evaluated TF-IDF and Count Vectorization together with multiple machine-learning classifiers and deep-learning models, illustrating the continued relevance of comparing conventional lexical approaches with contextual neural representations.

The applicability of traditional machine learning to Indonesian hoax detection has also been demonstrated in previous research. Nayoga et al. [6]investigated Indonesian hoax news using several deep neural network architectures together with traditional SVM and Naïve Bayes classifiers, with TF-IDF used to represent text for the conventional classifiers. Other Indonesian studies summarized in more recent literature have also reported competitive SVM performance. For example, a comparison of Naïve Bayes and SVM reported that SVM achieved higher classification accuracy, while another experiment comparing TF-IDF and IndoBERT-based representations within a BiLSTM framework obtained its strongest result using TF-IDF[1]. These findings suggest that lexical representations can remain effective for Indonesian fake-news classification despite the increasing adoption of contextual language models.

More recent research has increasingly focused on transformer-based models because they can represent words according to their surrounding context. IndoBERT is particularly relevant because it is pretrained for

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115–126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

Indonesian-language processing and can capture contextual relationships that cannot be represented explicitly by frequency-based vectors. Pekandi et al. [1] compared IndoBERT with a traditional ensemble consisting of Naïve Bayes, [7]Random Forest, [8]and SVM, as well as a CNN-LSTM model, using 6,120 equally balanced Indonesian news titles. IndoBERT achieved 92.24% accuracy and 92.23% F1-score, while the traditional ensemble achieved 91.18% accuracy. The study therefore demonstrated the potential advantage of contextual representations for Indonesian hoax classification, although it also noted that IndoBERT requires greater computational resources and longer training time than traditional approaches [1].

Research in other Indonesian misinformation settings further supports the potential of contextual models. Prastyapradipta et al. [3] investigated deep-learning and NLP approaches for Indonesian political hoax detection and highlighted the growing use of BERT-based methods for capturing Indonesian linguistic characteristics. Cahyo et al. [4], although focusing on Named Entity Recognition rather than binary hoax classification, found that a BERT-based model achieved the highest F1-score of 87.38% when identifying entities in Indonesian fake-news texts from the presidential election period. These findings provide additional evidence that contextual BERT-based representations are effective for processing linguistic information contained in Indonesian misinformation, although performance across tasks and datasets cannot be assumed to be identical.

The effectiveness of transformer models also depends on the composition and quantity of training data. Kusumawardani et al. [9] evaluated IndoBERT and Naïve Bayes on an Indonesian hoax dataset containing 600 headlines and investigated different training-test proportions together with back-translation augmentation. Their results showed that IndoBERT performance increased when additional training data and augmented examples were available, with the best configuration achieving an F1-score of 0.84 and accuracy of 0.87. Similarly, transformer-based fake-news studies outside the Indonesian context have reported strong results. Rahman et al. [9] compared multiple conventional classifiers, ensemble approaches, CNN, and BERT, with BERT achieving the strongest accuracy in their experimental setting. Amandeep and Suresh [10] also reported strong fake-news classification results using DistilBERT, demonstrating that pretrained transformer representations can provide high predictive performance while reducing some of the computational cost associated with full BERT architectures.

Nevertheless, greater model complexity does not guarantee superior performance in every experimental condition. Embarak [11] compared an Artificial Neural Network with traditional baselines including SVM and showed that deep learning could outperform conventional classifiers, but also emphasized that advanced models require more computational resources and sufficiently large labeled datasets. Zhou and Cai [5] similarly observed that deep-learning approaches offer richer semantic extraction than traditional TF-IDF-based models but introduce additional challenges, including computational requirements and class imbalance. Thus, reported superiority of deep or transformer-based models needs to be interpreted in relation to dataset size, representation strategy, class distribution, and experimental protocol rather than model architecture alone.

Another important issue is how news content is represented. Many fake-news studies focus primarily on headlines or treat text components independently. Alghamdi et al. [12] argued that the semantic relationship between a news headline and its associated body contains useful information for fake-news detection and that these relationships have not always been fully exploited. Headlines may attract readers or summarize a news item, but misleading headlines may not accurately represent the accompanying content. Modeling both title and body information can therefore provide a broader representation than headline-only classification. This consideration is relevant to Indonesian hoax detection because several previous studies have used news titles or headlines as the primary textual input [1], [9].

Class imbalance introduces an additional methodological challenge. A classifier trained on an unequal class distribution may obtain relatively high overall accuracy while still performing poorly on the minority class. Zhou and Cai [5] explicitly identified class imbalance as an important issue in fake-news datasets and used Macro F1-score to provide a more balanced assessment across classes. Malik et al. [8] similarly incorporated focal loss into an ensemble Graph Neural Network framework to mitigate class imbalance and improve learning from harder-to-classify observations. Their study demonstrates that imbalance-handling strategies can affect model behavior and that evaluation should consider performance beyond aggregate accuracy. This issue is especially important when comparing models on naturally imbalanced data rather than artificially balancing the distribution before evaluation.

Previous studies have also investigated increasingly complex fake-news architectures. Malik et al. [13] combined user-engagement patterns, BERT text embeddings, multiple graph neural networks, and ensemble strategies, while Alghamdi et al. [12] modeled semantic relationships among news content, headlines, and user information. Other work has incorporated attention, contextual embeddings, and stylistic features to improve fake-news classification. Such developments demonstrate that contemporary fake-news research is moving toward richer representations and more complex architectures. However, they also reinforce the need to determine whether additional representational complexity consistently produces meaningful predictive gains over strong and computationally simpler baselines.[14], [15].

Based on the reviewed literature, an important research gap remains regarding the comparative behavior of conventional lexical and contextual transformer representations for Indonesian hoax detection under identical and naturally imbalanced experimental conditions. Previous Indonesian studies have demonstrated promising performance using both TF-IDF-based conventional classifiers and BERT-based models; however, their experimental settings vary considerably in terms of dataset size, class distribution, input representation, balancing

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

strategy, and model-selection procedure. Some studies employ balanced datasets, resampling or augmentation, relatively small datasets, or headline-only inputs, making direct conclusions regarding the superiority of contextual models difficult. This limitation is particularly relevant for Indonesian hoax detection because naturally occurring datasets may contain substantial class imbalance, causing aggregate accuracy to be dominated by the majority class and potentially obscuring minority-class performance. [16].

Furthermore, the expected advantage of contextual language models cannot be assumed to hold uniformly across Indonesian hoax datasets. Although IndoBERT can capture contextual relationships that are not explicitly represented by frequency-based features, TF-IDF combined with SVM remains effective in high-dimensional sparse text spaces and can exploit discriminative lexical patterns with relatively low computational complexity. This creates an important empirical question: whether the additional contextual modeling capability of IndoBERT provides a meaningful performance advantage over a strong TF-IDF–SVM baseline when both models are trained and evaluated using the same data partitions, textual inputs, and model-selection protocol under substantial class imbalance. Addressing this question is important not only for predictive performance but also for determining whether increased model complexity is justified for this classification setting. [17].

Therefore, this study comparatively evaluates TF-IDF–SVM and IndoBERT for Indonesian hoax news classification using combined news titles and narratives under the same training, validation, and held-out test partitions. Unlike comparisons that rely primarily on overall accuracy or heterogeneous experimental settings, this study preserves the naturally imbalanced class distribution and uses Macro F1-score as the primary model-selection criterion to provide equal consideration to both classes. The study further examines the effect of class weighting on model behavior and evaluates accuracy, macro precision, macro recall, Macro F1-score, Weighted F1-score, and class-level performance. Finally, the predictions of the two selected models on the same held-out test observations are compared using the exact McNemar test to determine whether their difference in classification correctness is statistically significant.

The main contribution of this study is therefore not merely a comparison between two classification algorithms, but a controlled empirical evaluation of two fundamentally different text-representation paradigms—sparse lexical representation and contextual pretrained language modeling—under an identical and naturally imbalanced Indonesian hoax detection setting. By combining validation-based model selection, class-sensitive evaluation, and paired statistical testing, this study provides evidence on whether the increased representational complexity of IndoBERT translates into a meaningful predictive advantage over the computationally simpler TF-IDF–SVM baseline under the investigated conditions.

2. RESEARCH METHODOLOGY

This study employed a controlled comparative experimental design to evaluate two text-classification paradigms for Indonesian hoax detection: a conventional machine-learning approach based on Term Frequency–Inverse Document Frequency with Support Vector Machine (TF-IDF–SVM) and a contextual pretrained language model based on IndoBERT. Both approaches were developed and evaluated using identical training, validation, and held-out test partitions to ensure that performance differences were attributable to the modeling approaches rather than differences in data allocation. Particular attention was given to the naturally imbalanced class distribution, validation-based model selection, class-sensitive evaluation metrics, and paired statistical comparison of the final predictions. [18]

2.1 Research Stages

The research stages consisted of dataset preparation, data auditing, text preprocessing, stratified data splitting, model development, validation-based model selection, final testing, and statistical analysis. The dataset was first examined for missing values, duplicate records, and class distribution. After data cleaning, the title and narrative fields were combined to form the textual input. The resulting dataset was then divided using stratified sampling into 70% training, 15% validation, and 15% testing subsets, ensuring that the original class proportions were approximately preserved across all partitions. The same partitions were used for TF-IDF–SVM and IndoBERT to provide a controlled comparison.

Model development and configuration selection were conducted using only the training and validation data. Macro F1-score on the validation set was used as the primary selection criterion because of the substantial class imbalance. The held-out test set was not used for hyperparameter selection and was accessed only after the final configuration of each model had been determined. This separation was intended to reduce information leakage and provide a more objective estimate of generalization performance. The complete research workflow is illustrated in Figure 1.

Figure 1, the research workflow consists of several systematic stages, beginning with dataset preparation and ending with statistical analysis. First, the dataset is subjected to a data audit to assess data quality, identify duplicate or missing records, and examine the distribution of class labels. The text data are then preprocessed to obtain a clean and consistent representation suitable for model development. Subsequently, the dataset is divided using a stratified splitting strategy to maintain the class distribution across the training, validation, and test sets. Two classification

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

approaches are implemented and compared: a conventional machine learning approach using TF-IDF features with a Support Vector Machine (SVM) classifier and a transformer-based approach using IndoBERT. The validation set is used for model selection and hyperparameter optimization, while the final test set is reserved for an unbiased evaluation of model performance on previously unseen data. Finally, the performance of both approaches is evaluated using appropriate classification metrics and statistical analysis to determine whether the observed differences between the models are statistically significant.

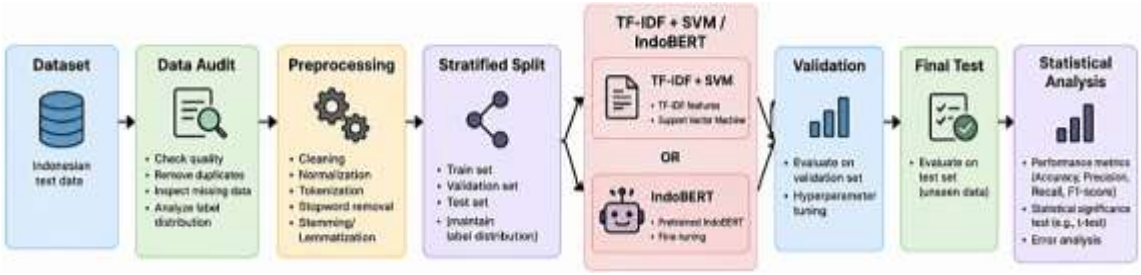

Figure 1. Research workflow for comparing TF-IDF–SVM and IndoBERT models

2.2 Dataset and Preprocessing

The Indonesia False News (Hoax) dataset was used in this study. The original labeled dataset contained 4,231 records consisting of ID, label, date, title, narrative, and image filename. The title and narrative fields were used as the textual information for classification. A preliminary data audit was conducted to identify missing values, duplicate records, and class-distribution characteristics. No missing values were identified in the variables used for classification.

Duplicate checking was performed after light normalization of the title and narrative fields. A record was considered duplicated only when the combined normalized title and narrative were identical; duplicate titles or narratives were not independently removed. Three duplicate copies were identified and removed, resulting in 4,228 unique observations. The resulting dataset contained 3,463 Label 1 samples (81.91%) and 765 Label 0 samples (18.09%), confirming a substantial class imbalance. Rather than artificially balancing the complete dataset before splitting, the natural distribution was retained to evaluate both approaches under the original imbalanced condition.

The cleaned dataset was divided using stratified sampling with a fixed random state of 42 into 2,959 training samples, 634 validation samples, and 635 held-out test samples, corresponding approximately to a 70:15:15 ratio. Stratification was applied to preserve the class distribution across the three subsets. Identical partitions were used for both modeling approaches. [1].

Different preprocessing procedures were applied according to the representation requirements of each model. For TF-IDF–SVM, the text was converted to lowercase and cleaned by removing URLs, HTML tags, special characters, and redundant whitespace. The cleaned title and narrative were combined and transformed into TF-IDF unigram and bigram representations with a maximum of 50,000 features. For IndoBERT, the combined title and narrative were processed using the tokenizer associated with the pretrained IndoBERT model. The maximum input sequence length was set to 128 tokens. This model-specific preprocessing preserved the distinction between sparse lexical representation in TF-IDF and contextual token representation in IndoBERT.

2.3 Model Development and Evaluation

The first classification approach combined TF-IDF representation with Linear Support Vector Classification (LinearSVC). Several SVM configurations were evaluated using the validation set by varying the regularization parameter and class-weighting strategy. The evaluated regularization values were C = 0.1, 0.5, and 1.0, with each configuration tested using either balanced class weighting or no class weighting. Macro F1-score was used as the primary model-selection criterion rather than accuracy because the dataset exhibited substantial class imbalance. The configuration producing the highest validation Macro F1-score was selected for final testing. This procedure resulted in a balanced LinearSVC with C = 0.1 as the final SVM configuration[5]..

The second approach employed the pretrained indobenchmark/indobert-base-p1 model for sequence classification. IndoBERT was selected because it provides contextual representations pretrained specifically for Indonesian-language processing and therefore represents an appropriate contextual counterpart to the sparse lexical TF-IDF representation. The combined title and narrative were tokenized using the corresponding IndoBERT tokenizer with a maximum sequence length of 128 tokens. The model was fine-tuned using a learning rate of 2×10^{-5}, a batch size of 16, and a maximum of five epochs. Weighted and unweighted cross-entropy configurations were evaluated to examine whether explicit class weighting improved performance under the imbalanced distribution. The final IndoBERT configuration was selected exclusively according to validation Macro F1-score, and the unweighted configuration was retained because it achieved the stronger validation Macro F1-score.

To make the experiment reproducible while reducing textual explanation, the final model configurations are summarized in Table 1.

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

Table 1. Experimental configuration

Parameter	TF-IDF + SVM	IndoBERT
Input	Title + narrative	Title + narrative
Text representation	TF-IDF unigram–bigram	IndoBERT tokenizer
Maximum features/length	50,000 features	128 tokens
Main classifier	LinearSVC	IndoBERT sequence classifier
Main parameter	C = 0.1	LR = 2×10^{-5}
Class weighting	Balanced	Unweighted
Batch size	–	16
Maximum epoch	–	5
Selection metric	Macro F1	Macro F1

Performance was measured using accuracy, macro precision, macro recall, Macro F1-score, Weighted F1-score, and confusion matrices. Macro F1 was used as the primary metric because it gives equal consideration to both classes and is more informative for imbalanced classification [19]. Previous fake news research has similarly used Macro F1 when evaluating imbalanced datasets.

The hyperparameter search was intentionally limited to a controlled set of configurations rather than an exhaustive optimization of either model. The objective of the study was to compare representative conventional and contextual approaches under the same experimental protocol rather than to establish the maximum achievable performance of each architecture. Accordingly, the results should be interpreted within the investigated configuration space. In particular, the use of a single IndoBERT variant, a fixed maximum sequence length of 128 tokens, and a limited fine-tuning configuration may constrain the extent to which the transformer model's full potential is represented. Future experiments may extend this comparison by evaluating additional IndoBERT variants, sequence lengths, learning rates, training epochs, and parameter-efficient fine-tuning strategies.

To reduce test-set bias, all model-selection decisions were based exclusively on validation performance, and the held-out test set was used only once for the final evaluation of the selected configurations. Nevertheless, repeated comparison of multiple configurations on a single validation partition can introduce a degree of selection bias toward that validation set. Because nested cross-validation was not employed in the present study, this possibility cannot be completely excluded. Therefore, the reported test results are interpreted as performance estimates for the selected configurations rather than as evidence of globally optimal hyperparameters. Future studies may employ repeated stratified cross-validation or nested cross-validation to obtain more robust model-selection estimates.

Final model performance was evaluated on the same held-out test set of 635 observations using accuracy, macro precision, macro recall, Macro F1-score, Weighted F1-score, and confusion matrices. Macro F1-score was emphasized because it assigns equal importance to both classes and is therefore more informative than overall accuracy when evaluating imbalanced classification. Class-level precision, recall, and F1-score were also examined to identify differences in majority- and minority-class behavior.

In addition to descriptive performance comparison, the final predictions generated by TF-IDF–SVM and IndoBERT for the same 635 test observations were compared using the exact McNemar test. This paired statistical test evaluates whether the two classifiers differ significantly in their correctness patterns on identical observations. The null hypothesis (H_0) states that the two models have equivalent probabilities of correct classification, whereas the alternative hypothesis (H_1) states that their paired correctness differs. Statistical significance was assessed at $\alpha = 0.05$. A p-value below 0.05 was considered evidence for rejecting H_0, whereas a p-value equal to or greater than 0.05 indicated insufficient evidence to conclude a statistically significant difference between the classifiers. [20].

3. RESULT AND DISCUSSION

This section presents and discusses the experimental results of TF-IDF–SVM and IndoBERT for Indonesian hoax detection. The analysis covers validation-based model selection, final held-out test performance, class-level classification behavior, the effect of class weighting, and paired statistical comparison using the exact McNemar test. Particular attention is given to whether the numerical performance differences between the two approaches represent statistically meaningful differences and to how the naturally imbalanced class distribution may have influenced their behavior.

3.1 SVM Model Selection and Performance

The SVM configuration was determined using the validation set before evaluating the model on the test set. Six configurations were evaluated by combining three regularization values ($C = 0.1$, 0.5, and 1.0) with either no class weighting or balanced class weighting. Macro F1-score was used as the primary selection criterion because the

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

dataset contained an unequal distribution between Label 0 and Label 1. The use of Macro F1 is particularly relevant for imbalanced fake news classification because it assigns equal importance to the performance of each class [5].

Table 2. SVM validation results

C	Class Weight	Accuracy	Macro F1
0.1	Balanced	0.8265	0.7018
0.5	Balanced	0.8328	0.6952
1.0	Balanced	0.8328	0.6775
1.0	None	0.8549	0.6561
0.5	None	0.8502	0.6301
0.1	None	0.8297	0.5103

As shown in Table 2, the highest validation accuracy was obtained by the unweighted SVM configuration with C = 1.0, whereas the highest Macro F1-score of 0.7018 was obtained using balanced class weighting with C = 0.1. Consequently, the latter configuration was selected for final evaluation. This difference illustrates an important consequence of class imbalance: optimizing overall accuracy does not necessarily produce the most balanced performance across classes. Because Label 1 substantially dominated the dataset, a model could achieve relatively high accuracy by performing strongly on the majority class while providing considerably weaker discrimination for Label 0. The use of Macro F1 as the selection criterion therefore reduced the influence of majority-class dominance on model selection.[5].

The selected SVM model was subsequently evaluated on the held-out test set. Its class-level performance is presented in Table 3.

Table 3. Final SVM test performance

Class	Precision	Recall	F1-score	Support
Label 0	0.5315	0.5130	0.5221	115
Label 1	0.8931	0.9000	0.8966	520
Macro Avg.	0.7123	0.7065	0.7093	635
Weighted Avg.	—	—	0.8287	635

On the held-out test set, the selected TF-IDF–SVM model achieved an accuracy of 0.8299 and a Macro F1-score of 0.7093. However, class-level analysis reveals a substantial difference between the two labels. Label 1 achieved an F1-score of 0.8966, whereas Label 0 achieved only 0.5221. Thus, although balanced class weighting improved the validation criterion used for model selection, it did not eliminate the difficulty associated with the minority class. This finding reinforces the importance of reporting class-level and macro-averaged metrics rather than relying solely on accuracy when evaluating naturally imbalanced hoax datasets..

The competitiveness of SVM is consistent with earlier work showing that traditional classifiers remain useful for fake news detection. Nayoga et al. [6] applied SVM and Naïve Bayes alongside multiple deep neural networks to Indonesian hoax news classification, while other Indonesian studies summarized by Pekandi et al. reported that SVM could outperform Naïve Bayes [1].

3.2 IndoBERT Model Selection and Performance

For IndoBERT, two loss configurations were compared during validation: weighted cross-entropy and standard unweighted cross-entropy. This experiment was conducted to determine whether explicit compensation for the imbalanced class distribution improved overall classification balance.

Table 4. IndoBERT validation results

Loss Configuration	Accuracy	Macro Precision	Macro Recall	Macro F1
Weighted CE	0.7240	0.6428	0.7197	0.6477
Unweighted CE	0.8123	0.6745	0.6484	0.6591

The IndoBERT results indicate that contextual representation alone did not guarantee superior classification performance under the investigated conditions. Although IndoBERT is designed to capture semantic relationships among tokens within their surrounding context, its performance remained affected by the characteristics of the training data and the selected fine-tuning configuration. In particular, the substantial class imbalance may have limited the number and diversity of minority-class examples available for learning discriminative contextual patterns. Furthermore, the use of a maximum sequence length of 128 tokens may truncate information from longer combined title–narrative inputs, although the present experiment did not explicitly measure the proportion or content of truncated sequences. Therefore, sequence truncation should be regarded as a plausible limitation rather than a demonstrated cause of the observed performance difference.

This finding illustrates that class weighting does not necessarily improve every evaluation measure simultaneously. Previous fake news studies have identified uneven class distributions as an important classification

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

problem and proposed loss-based mechanisms such as focal loss to increase attention to difficult or underrepresented samples [5]. The present results indicate that the effectiveness of such imbalance-handling strategies remains dependent on the characteristics of the dataset and the model-selection objective.

After the loss configuration was selected using validation data, the unweighted IndoBERT model was evaluated on the held-out test set.

Table 5. Final IndoBERT test performance

Class	Precision	Recall	F1-score	Support
Label 0	0.4579	0.4261	0.4414	115
Label 1	0.8750	0.8885	0.8817	520
Macro Avg.	0.6665	0.6573	0.6616	635
Weighted Avg.	—	—	0.8020	635

As shown in Table 5, IndoBERT achieved an accuracy of 0.8047 and Macro F1-score of 0.6616. Similar to SVM, performance was considerably stronger for Label 1 than Label 0. The F1-score reached 0.8817 for Label 1 but only 0.4414 for Label 0. The validation and test Macro F1-scores were 0.6591 and 0.6616, respectively, indicating that the selected model maintained similar performance on unseen test data.

The comparison between weighted and unweighted IndoBERT configurations also demonstrates a trade-off in addressing class imbalance. Class weighting can increase the model's sensitivity to minority-class observations, but greater minority recall does not necessarily translate into a higher Macro F1-score when accompanied by reduced precision or weaker majority-class performance. The unweighted configuration was therefore retained because it produced the stronger validation Macro F1-score. This finding suggests that simple loss reweighting alone may not be sufficient to resolve the imbalance problem in this dataset.

3.3 Final Comparison of SVM and IndoBERT

The primary comparison between the two approaches is summarized in Table 6. All values in this table were obtained from the same held-out test set containing 635 observations.

Table 6. Final performance comparison

Metric	TF-IDF + SVM	IndoBERT	Difference
Accuracy	0.8299	0.8047	+0.0252
Macro Precision	0.7123	0.6665	+0.0458
Macro Recall	0.7065	0.6573	+0.0492
Macro F1	0.7093	0.6616	+0.0477
Weighted F1	0.8287	0.8020	+0.0267

As shown in Table 6, The final comparison demonstrates a consistent numerical advantage for TF-IDF–SVM across the aggregate evaluation metrics. TF-IDF–SVM achieved an accuracy of 0.8299, compared with 0.8047 for IndoBERT, while the corresponding Macro F1-scores were 0.7093 and 0.6616, respectively. Thus, the conventional model achieved a Macro F1 advantage of approximately 0.0477. These results indicate that, within the investigated experimental setting, increasing representational complexity from sparse lexical features to contextual transformer representations did not translate into higher predictive performance.

This result is particularly noteworthy when compared with previous Indonesian hoax detection studies. Pekandi et al. [1] evaluated an ensemble of Naïve Bayes, Random Forest, and SVM, CNN-LSTM, and IndoBERT on 6,120 equally balanced Indonesian news titles. IndoBERT achieved 92.24% accuracy, slightly exceeding the traditional ensemble at 91.18% [1]. In contrast, the present experiment used 4,228 samples with an approximately 81.91%–18.09% class distribution, combined titles and narratives, and found SVM to produce higher test Macro F1 than IndoBERT.

One plausible explanation is that discriminative lexical patterns remain highly informative for the present hoax dataset. TF-IDF directly emphasizes terms and n-grams that differ in importance across documents, while a linear SVM can exploit such high-dimensional sparse representations effectively. If particular words, expressions, or short lexical combinations are strongly associated with the class labels, the conventional approach may construct an effective decision boundary without requiring deeper contextual representation. This interpretation is consistent with the observed performance but should not be interpreted as evidence that lexical representation is universally superior to contextual modeling.

Furthermore, Pekandi et al. [1] reported that IndoBERT requires greater computational resources and longer training time, while traditional methods are computationally lighter [1] .Therefore, the present finding has a practical implication: when the performance advantage of a transformer model is not guaranteed for a particular dataset, a TF-IDF + SVM pipeline may remain an effective alternative with substantially lower model complexity.

A related Indonesian study reported by Pekandi et al. also found that a Bi-LSTM experiment comparing TF-IDF and IndoBERT-based representations achieved its best result using TF-IDF, with 92% accuracy [1] **.**This

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

provides additional evidence that contextual embeddings do not automatically guarantee superior results for every Indonesian fake news dataset.

Kusumawardani et al. [9] provide another useful comparison. Their Indonesian hoax dataset contained only 600 headlines, consisting of 372 valid and 228 hoax samples, and their study examined Naïve Bayes and IndoBERT under different split ratios and back-translation augmentation strategies . Although the experimental setting differs from the present work, the study further illustrates that Indonesian hoax detection performance is sensitive to dataset size, training composition, and augmentation strategy. Therefore, direct comparison of raw accuracy values across studies should be performed cautiously.

Despite the numerical advantage of TF-IDF–SVM, the exact McNemar test provides an important qualification to the descriptive comparison. The test produced a p-value of 0.1371, which is greater than the predefined significance level of $\alpha = 0.05$. Therefore, the null hypothesis of equivalent paired classification correctness cannot be rejected. In other words, although TF-IDF–SVM achieved higher accuracy and Macro F1-score, the pattern of disagreements between the two classifiers on the same held-out test observations does not provide sufficient statistical evidence to conclude that their classification correctness differs significantly.

This distinction between numerical and statistical superiority is central to interpreting the experiment. Aggregate metrics summarize the magnitude of observed performance, whereas the McNemar test examines whether the classifiers exhibit significantly different correctness patterns on paired observations. The present results therefore support the conclusion that TF-IDF–SVM was numerically stronger under the investigated setting, but they do not support a claim that SVM is statistically superior to IndoBERT. Consequently, the results should be described in terms of a competitive conventional baseline rather than definitive superiority of one modeling paradigm.

3.4 Confusion Matrix and Error Analysis

The confusion matrices provide additional insight into the behavior of both classifiers beyond aggregate metrics.

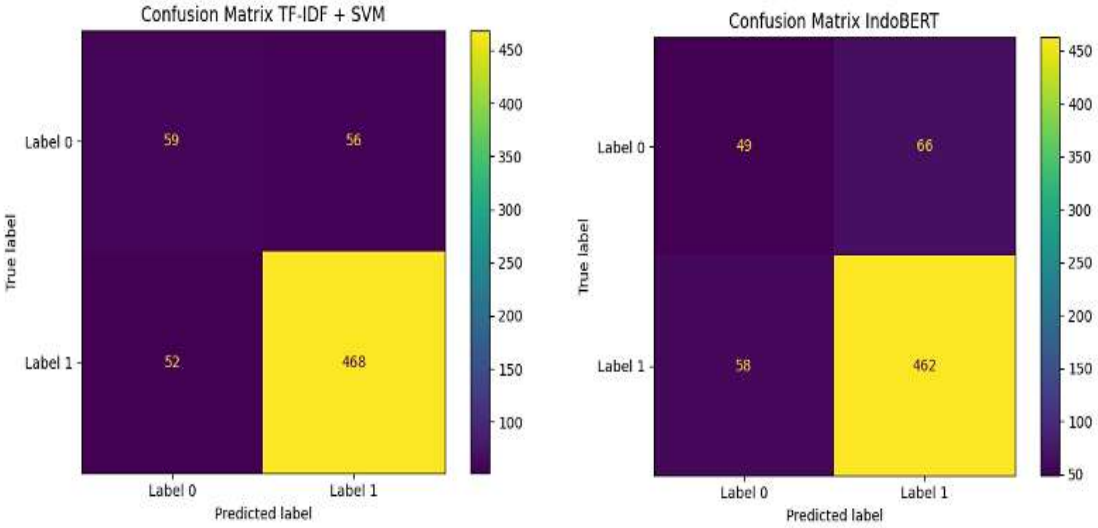

Figure 2. Confusion matrices of TF-IDF + SVM and IndoBERT on the test set

As illustrated in Figure 2, TF-IDF + SVM correctly classified 59 of 115 Label 0 samples and 468 of 520 Label 1 samples, whereas IndoBERT correctly classified 49 Label 0 samples and 462 Label 1 samples. The SVM model therefore produced fewer classification errors for both classes. The difference was more apparent for Label 0, for which SVM generated 56 misclassifications compared with 66 produced by IndoBERT. These results indicate that the minority class remained more difficult for both classifiers, although SVM showed relatively better discrimination under the imbalanced test distribution

3.5 Statistical Comparison Using McNemar Test

Although SVM obtained higher values for all final aggregate metrics, a numerical difference does not necessarily indicate that the two classifiers have statistically different error behavior. Therefore, an exact McNemar test was performed using their paired predictions on the same 635 test observations.

Table 7. Paired prediction outcomes for McNemar analysis

Prediction Outcome	Number of Samples
Both models correct	468
SVM correct, IndoBERT incorrect	59
SVM incorrect, IndoBERT correct	43

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

Both models incorrect	65

As shown in Table 7, both classifiers correctly predicted 468 test samples and incorrectly predicted the same 65 samples. However, their errors were not completely identical. SVM correctly classified 59 observations that IndoBERT misclassified, while IndoBERT correctly classified 43 observations that SVM misclassified. Thus, there were 102 observations for which the models produced different correctness outcomes.

The exact McNemar test produced a statistic of 43.0 and a p-value of 0.1371. With a significance level of $\alpha = 0.05$, the null hypothesis cannot be rejected because $p > 0.05$. Therefore, although SVM achieved higher accuracy and Macro F1 numerically, the paired correctness difference between SVM and IndoBERT was not statistically significant at the 5% level**.**

McNemar's test has been used in NLP classification studies to examine paired misclassification distributions when models are applied to the same observations [20]. Importantly, the McNemar result should not be interpreted as a direct statistical test of the difference in Macro F1-score. Instead, it evaluates the discordance in paired classification correctness.

This statistical result provides a more conservative interpretation of the performance comparison. The descriptive metrics favor SVM, but the evidence is insufficient to conclude that SVM has a statistically significant advantage in paired prediction correctness. Moreover, the 59 SVM-only correct predictions and 43 IndoBERT-only correct predictions indicate that the models capture partially different decision patterns. This complementarity may

3.6 Discussion

The experimental results demonstrate that TF-IDF–SVM remained competitive with IndoBERT for Indonesian hoax detection under the naturally imbalanced data distribution investigated in this study. TF-IDF–SVM achieved an accuracy of 0.8299 and a Macro F1-score of 0.7093, whereas IndoBERT obtained an accuracy of 0.8047 and a Macro F1-score of 0.6616. Although the conventional approach therefore produced higher descriptive performance, this result differs from several previous studies in which transformer-based representations achieved stronger classification performance. The difference indicates that model performance should be interpreted in relation to dataset characteristics, class distribution, input representation, and experimental configuration rather than model complexity alone.

The present findings differ particularly from those reported by Pekandi et al. [1] . Their study compared IndoBERT with traditional machine-learning and deep-learning approaches using 6,120 equally balanced Indonesian news titles. IndoBERT achieved an accuracy of 92.24% and an F1-score of 92.23%, slightly outperforming the traditional ensemble, which achieved 91.18% accuracy. In contrast, the present study retained the naturally imbalanced distribution of 4,228 observations, consisting of 81.91% Label 1 and 18.09% Label 0, and used combined titles and narratives rather than titles alone. Under these conditions, TF-IDF–SVM produced a higher Macro F1-score than IndoBERT. These contrasting findings suggest that the advantage of contextual representations observed on balanced data cannot automatically be generalized to naturally imbalanced Indonesian hoax datasets.

The competitiveness of the conventional approach is also consistent with previous Indonesian fake-news research. Nayoga et al. [6] evaluated SVM and Naïve Bayes together with several deep neural network architectures for Indonesian hoax classification and demonstrated that conventional text-classification approaches remain relevant. Moreover, previous experiments summarized by Pekandi et al. [1] reported that SVM could outperform Naïve Bayes and that a Bi-LSTM experiment comparing TF-IDF and IndoBERT-based representations obtained its strongest result using TF-IDF, reaching approximately 92% accuracy. These findings support the present observation that contextual embeddings do not necessarily guarantee better predictive performance when discriminative lexical patterns provide sufficient information for classification.

The results can also be interpreted in relation to the findings of Kusumawardani et al. [9]. who evaluated IndoBERT and Naïve Bayes using an Indonesian hoax dataset containing 600 headlines. Their results showed that IndoBERT performance improved when additional training examples were introduced through different training proportions and back-translation augmentation, with the best configuration achieving an F1-score of 0.84 and accuracy of 0.87. This finding suggests that transformer performance is sensitive to the amount and composition of training data. In the present study, the minority class accounted for only 18.09% of the observations, potentially limiting the diversity of minority-class contextual patterns available during fine-tuning. Thus, the weaker IndoBERT performance observed here may partly reflect the interaction between contextual modeling and the naturally imbalanced training distribution rather than an inherent weakness of IndoBERT itself.

The importance of class imbalance is further consistent with Zhou and Cai [5] , who identified uneven class distributions as an important challenge in fake-news classification and emphasized the use of Macro F1-score for evaluating performance across classes. Malik et al, [8] similarly employed focal-loss-based mechanisms to increase attention to difficult or underrepresented observations. In the present experiment, class weighting also affected model behavior. Balanced weighting produced the highest validation Macro F1-score for TF-IDF–SVM, whereas weighted cross-entropy for IndoBERT increased sensitivity to the minority class but did not produce the highest overall Macro F1-score. These results indicate that imbalance-handling strategies involve a trade-off between minority-class sensitivity and overall classification balance and that no single weighting mechanism can be assumed to improve all performance measures simultaneously.

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

The present findings are also consistent with the broader observation of Embarak [11] that advanced neural models can provide strong performance but generally require larger labeled datasets and greater computational resources. Similarly, Zhou and Cai [5] noted that deep-learning approaches provide richer semantic representations than traditional TF-IDF-based approaches but introduce additional computational and data-related requirements. Therefore, the numerical advantage obtained by TF-IDF–SVM in this study has practical relevance. When discriminative lexical patterns are strong, the labeled dataset is relatively limited or imbalanced, and computational efficiency is an important consideration, a conventional TF-IDF–SVM pipeline may remain a reasonable alternative to a substantially more complex transformer model.

Several experimental factors may additionally explain why IndoBERT did not reproduce the superiority reported in some previous studies. The present IndoBERT configuration used a maximum sequence length of 128 tokens for the combined title and narrative input. Consequently, information contained beyond this limit may not have contributed to contextual representation for longer documents. Furthermore, only indobenchmark/indobert-base-p1 and a limited fine-tuning configuration were evaluated. These factors should not be interpreted as confirmed causes because this study did not perform dedicated ablation experiments involving different sequence lengths, tokenization strategies, IndoBERT variants, or broader hyperparameter configurations. Instead, they identify experimental conditions that may be investigated in future research to determine when contextual modeling provides a meaningful advantage over lexical representations.

Most importantly, the descriptive comparison must be interpreted together with the exact McNemar test. Although TF-IDF–SVM achieved higher accuracy, Macro Precision, Macro Recall, Macro F1-score, and Weighted F1-score, the exact McNemar test produced $p = 0.1371$. Because this value is greater than $\alpha = 0.05$, there is insufficient statistical evidence to conclude that TF-IDF–SVM and IndoBERT differ significantly in paired classification correctness. The 59 observations correctly classified only by SVM and the 43 observations correctly classified only by IndoBERT further indicate that the two approaches capture partially different decision patterns. Therefore, the principal finding of this study is not that TF-IDF–SVM is universally superior to IndoBERT, but that a computationally simpler lexical model remains a competitive baseline under a naturally imbalanced Indonesian hoax-detection setting. This conclusion complements previous studies reporting stronger IndoBERT performance by demonstrating that the benefit of contextual modeling is dependent on dataset composition, class balance, textual input, and fine-tuning conditions.

4. CONCLUSION

This study comparatively evaluated TF-IDF–SVM and IndoBERT for Indonesian hoax detection under an identical experimental protocol using a naturally imbalanced dataset. The results showed that TF-IDF–SVM achieved higher numerical performance, with an accuracy of 0.8299 and a Macro F1-score of 0.7093, compared with 0.8047 and 0.6616, respectively, for IndoBERT. However, the exact McNemar test yielded a p-value of 0.1371, indicating that the difference in paired classification correctness was not statistically significant at $\alpha = 0.05$. Therefore, the findings should not be interpreted as evidence of general statistical superiority of TF-IDF–SVM over IndoBERT. Instead, they demonstrate that a computationally simpler lexical approach can remain a competitive baseline against a contextual transformer model under the investigated imbalanced Indonesian hoax-detection setting. The results further indicate that model complexity alone does not guarantee improved classification performance, particularly when discriminative lexical patterns are informative and minority-class observations are limited. From a practical perspective, TF-IDF–SVM may be preferred when labeled data and computational resources are limited, class imbalance is substantial, and lexical cues provide strong discriminatory information. Transformer-based approaches such as IndoBERT remain promising when larger and more representative labeled datasets, richer contextual information, and sufficient computational resources are available for broader fine-tuning. Future research should evaluate additional Indonesian pretrained language models, longer input sequence lengths, broader fine-tuning configurations, advanced imbalance-handling strategies, and repeated or nested cross-validation to determine whether the observed performance pattern persists across different Indonesian hoax datasets and experimental conditions.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: Y.C and N;
Methodology: N and Y.C;
Software: Y.C and H.A;
Validation: Y.C, N, and H.A;
Formal Analysis: Y.C, N, and H.A;
Investigation: Y.C and H.A;
Resources: Y.C and C.S;

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

Data Curation: Y.C;
Writing Original Draft Preparation: Y.C;
Writing Review and Editing: Y.C, N, H.A, and C.S;
Visualization: Y.C and H.A.
All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The dataset used in this study is publicly available from the Indonesia False News (Hoax) dataset on Kaggle. The processed data and additional materials supporting the findings of this study are available from the corresponding author upon reasonable request.

5.3 Funding

This research received no external funding. The publication and manuscript processing fees associated with this article, if applicable, will be covered by the authors.

5.4 Institutional Review Board Statement

Not applicable. This study used a publicly available secondary dataset and did not involve human participants, clinical interventions, or the collection of personally identifiable information.

5.5 Informed Consent Statement

Not applicable. This study did not involve direct participation of human subjects or the collection of informed consent data.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used generative AI tools to assist in improving language clarity, grammatical accuracy, and the organization of the manuscript. The authors critically reviewed, verified, and edited all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the published work.

REFERENCES

[1] L. A. Pekandi, R. G. Widjaja, A. Ananta, J. Harefa, and K. Jingga, “Evaluating IndoBERT for Indonesian Hoax News Detection: A Comparative Study with Ensemble and CNN-LSTM Models,” in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 1625–1633. doi: 10.1016/j.procs.2025.09.105.

[2] K. S. Sreekar Datta, G. Narasimha Naidu, S. Abhishek, and T. Anjali, “Enhancing Veracity: Empirical Evaluation of Fake News Detection Techniques,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 97–107. doi: 10.1016/j.procs.2024.03.199.

[3] 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,” in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 1742–1751. doi: 10.1016/j.procs.2025.09.117.

[4] P. W. Cahyo, U. S. Aesyi, W. A. Setianto, and T. Sulaiman, “A Novel Named Entity Recognition approach of Indonesian fake news using part of speech and BERT model on presidential election,” Dec. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.jjimei.2025.100354.

[5] Q. H. Zhou and T. Cai, “Adaptive gate residual connection and multi-scale RCNN for fake news detection,” *Machine Learning with Applications*, vol. 19, Mar. 2025, doi: 10.1016/j.mlwa.2024.100612.

[6] B. P. Nayoga, R. Adipradana, R. Suryadi, and D. Suhartono, “Hoax Analyzer for Indonesian News Using Deep Learning Models,” in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 704–712. doi: 10.1016/j.procs.2021.01.059.

[7] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, “Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 349–356. doi: 10.1016/j.procs.2024.03.014.

[8] S. T. Hamidou and A. Mehdi, “Enhancing IDS performance through a comparative analysis of Random Forest, XGBoost, and Deep Neural Networks,” *Machine Learning with Applications*, vol. 22, p. 100738, Dec. 2025, doi: 10.1016/j.mlwa.2025.100738.

[9] R. Pradina Kusumawardani, M. R. Lukman, M. Ibrahim, A. Ilham, C. Waladsae Adiena, and R. Prayoga, “ScienceDirect The Effect of the Back Translation Data Augmentation Technique for Indonesian-Language Hoax News Detection Model,” *Procedia Comput. Sci.*, vol. 284, pp. 1479–1486, 2026, [Online]. Available: www.sciencedirect.com

[10] Amandeep and S. Suresh, “Transforming Fake News Detection: Leveraging DistilBERT Models for Enhanced Accuracy,” in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 283–290. doi: 10.1016/j.procs.2025.03.203.

[11] O. Embarak, “Deep Learning for Fake News Detection: Analysing Facebook’s Misinformation Networks,” in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 199–208. doi: 10.1016/j.procs.2025.07.173.

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 115−126
DOI: doi.org/10.64366/ijids.v3i2.575
https://journals.adaresearch.or.id/ijids/index

[12] J. Alghamdi, Y. Lin, and S. Luo, "Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media," *Eng. Appl. Artif. Intell.*, vol. 137, Nov. 2024, doi: 10.1016/j.engappai.2024.109240.
[13] A. Malik, D. K. Behera, J. Hota, and A. R. Swain, "Ensemble graph neural networks for fake news detection using user engagement and text features," *Results in Engineering*, vol. 24, Dec. 2024, doi: 10.1016/j.rineng.2024.103081.
[14] F. J. Lara-Abelenda, D. Chushig-Muzo, C. B. Acosta, A. M. Wägner, C. Granja, and C. Soguero-Ruiz, "Evaluating Time Series Classification Models for Nocturnal Hypoglycemia: From Predictive Performance to Environmental Impact," *IEEE Access*, vol. 13, no. September, pp. 150756–150771, 2025, doi: 10.1109/ACCESS.2025.3600917.
[15] A. Chahal *et al.*, "Predictive analytics technique based on hybrid sampling to manage unbalanced data in smart cities," *Heliyon*, vol. 10, no. 24, Dec. 2024, doi: 10.1016/j.heliyon.2024.e39275.
[16] M. A. Umar, Z. Chen, K. Shuaib, and Y. Liu, "Effects of feature selection and normalization on network intrusion detection," *Data Science and Management*, vol. 8, no. 1, pp. 23–39, Mar. 2025, doi: 10.1016/j.dsm.2024.08.001.
[17] W. Zhou and D. Yang, "Analysis and comparison of automatic image focusing algorithms in digital image processing," *J. Radiat. Res. Appl. Sci.*, vol. 16, no. 4, p. 100672, Dec. 2023, doi: 10.1016/j.jrras.2023.100672.
[18] R. Ranjan *et al.*, "A novel local optimal oriented pattern for image splicing detection leveraging deep learning and SVM," *Franklin Open*, vol. 16, Sep. 2026, doi: 10.1016/j.fraope.2026.100634.
[19] S. Nouas, L. Oukid, and F. Boumahdi, "Enhancing imbalanced text classification: an overlap-based refinement approach," *Data Science and Management*, vol. 8, no. 4, pp. 474–484, Dec. 2025, doi: 10.1016/j.dsm.2025.03.001.
[20] B. Widmer, G. Moffa, H. E. Viehweger, and M. Sangeux, "Use of natural language processing to predict the diagnoses of patients with gait impairments," *Inform. Med. Unlocked*, vol. 59, Jan. 2025, doi: 10.1016/j.imu.2025.101712.