Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1200~1209
e-ISSN: 2548-6861

Hybrid Lexical-Semantic Approach for Clickbait Classification in Indonesian News

Mohammad Nizar Farizi [1*], Lukmanul Hakim [2*], Sultan Ahmad Haidir [3*], Ema Utami[4*]
* Master of Informatics, Universitas AMIKOM Yogyakarta
1m.nizar@students.amikom.ac.id [1], l.hakim@students.amikom.ac.id [2], s.ahmad@students.amikom.ac.id [3]

Article Info

Article history:

Received 2026-01-18
Revised 2026-02-27
Accepted 2026-04-08

Keyword:

Clickbait,
Text Classification,
TF-IDF,
IndoBERT,
Hybrid Pre-Processing.

ABSTRACT

The rise of digital media in Indonesia has led to the proliferation of hyperbolic clickbait headlines that undermine media credibility and spread misinformation. This study proposes a hybrid lexical-semantic workflow for clickbait classification in Indonesian news by integrating lexical representations from Term Frequency-Inverse Document Frequency (TF-IDF) and token overlap ratios with deep contextual representations from IndoBERT embeddings. Using a dataset of 3,000 news articles from Detik.com labeled through a weak supervision strategy based on headline-content discrepancy, the extracted features were processed using a Single Layer Perceptron (SLP) and compared against Logistic Regression (LR) and Random Forest (RF) models. Evaluation via stratified 5-fold cross-validation focused on the F1-score to address class imbalance, revealing that hybrid features enhance model robustness against subtle misalignments. While LR reached a peak F1-score of 0.90 in lexical settings, the hybrid SLP configuration yielded a stable F1-score of 0.87, with feature importance analysis identifying IndoBERT semantic similarity as the most critical predictor. Ultimately, this hybrid approach successfully balances semantic depth and computational efficiency, offering an effective framework for safeguarding information integrity on Indonesian digital news platforms.

I. INTRODUCTION

The development of digital media in Indonesia has significantly transformed public news consumption patterns, accompanied by the widespread emergence of clickbait headlines [1]. These headlines are deliberately crafted to attract user attention through hyperbolic or misleading language, yet they often fail to accurately reflect the actual content of the article [2], [3]. Such practices not only diminish media credibility in the eyes of readers but also potentially exacerbate the dissemination of misinformation within digital spaces [1], [4]. Therefore, the development of automated clickbait detection systems supported by Natural Language Processing (NLP) technologies has become a critical necessity to safeguard information integrity on online news platforms [2].

In general, NLP approaches for clickbait detection can be categorized into two main types: lexical-based and semantic-based approaches [5]. Lexical approaches rely on surface-level textual analysis, such as word frequency and term occurrence patterns, employing widely used methods such as Term Frequency-Inverse Document Frequency (TF-IDF) [6]. Although lexical methods are recognized for their computational efficiency, they exhibit limited sensitivity to semantic variation and contextual ambiguity in Indonesian language usage, as they rely primarily on surface-level word frequency distributions rather than deeper contextual representations [1], [5].

Conversely, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) and its monolingual variant IndoBERT have demonstrated strong capability in modeling inter-sentential semantic relationships through contextualized representations [4]. Several studies conducted in Indonesia indicate that IndoBERT improves text classification accuracy by effectively capturing complex syntactic and semantic dependencies [1]. However, the standalone deployment of transformer models typically requires substantial computational resources and significantly longer training time compared to traditional machine learning models [5], [7].

These observations suggest a complementary potential between lexical and semantic representations, thereby motivating further exploration of integrated or hybrid approaches. Hybrid methods have been shown to enhance model robustness by combining surface-level lexical similarity features with deep contextual semantic representations [5]. Recent research demonstrated that integrating extractive algorithms such as TF-IDF with BERT (TF-IDF-BERT) yields significantly improved performance compared to using a standard BERT model alone in classification tasks [6]. This improvement arises because lexical methods remain relevant due to their efficiency and simplicity in capturing word frequency patterns, whereas transformer-based models excel at modeling contextual ambiguity in depth [5]. Nevertheless, systematic investigations examining lexical-semantic hybrid workflows within the context of Indonesian-language news, particularly concerning the relationship between headlines and article content, remain limited [5].

Recent advancements in Natural Language Processing (NLP) have shifted attention toward transformer-based architectures such as BERT and its Indonesian variant IndoBERT, which consistently outperform traditional approaches in capturing contextual meaning for clickbait classification [1], [4]. In practical implementation, researchers must determine whether to utilize transformer models as static feature extractors to generate embeddings or perform full fine-tuning to adapt model weights to domain-specific characteristics [4]. However, the application of transformer models to Indonesian news frequently encounters challenges related to enormous computational costs and the requirement for high-quality, large-scale datasets, which is particularly difficult for low-resource languages [1], [5].

Despite these technological advancements, systematic investigations that integrate lexical transparency with contextual semantic modeling specifically for detecting headline-content discrepancy in Indonesian news remain scarce [1], [5]. This gap underscores the necessity of a hybrid framework capable of balancing interpretability, computational feasibility, and semantic depth [5], [6].

Accordingly, this study proposes and analyzes a hybrid lexical-semantic preprocessing workflow based on TF-IDF and IndoBERT for clickbait classification in Indonesian news articles. The proposed workflow integrates two types of textual representations: lexical similarity (utilizing TF-IDF for word frequency analysis) and semantic similarity (utilizing IndoBERT embeddings to capture deep contextual relationships). The integrated representations are subsequently fed into a neural integration layer, implemented through a lightweight Single Layer Perceptron (SLP) model, to determine the final classification (Fakhruzzaman et al., 2021). Thus, this study focuses on examining the role and impact of hybrid preprocessing in improving the performance and efficiency of Indonesian clickbait classification systems.

II. METHODS

A. General Design of Research

This research developed a natural language processing (NLP)-based pipeline for the classification of clickbait in Indonesian-language news [5]. The design of the system consists of five main stages: online news data collection, text pre-processing, extraction of lexical and semantic features, integration of features in a neural integration layer, and model performance evaluation [2], [8]. The pipeline stages are shown in Figure 1, which depicts the workflow from raw data acquisition to final classification results, is designed to ensure computational feasibility while maintaining high detection accuracy.

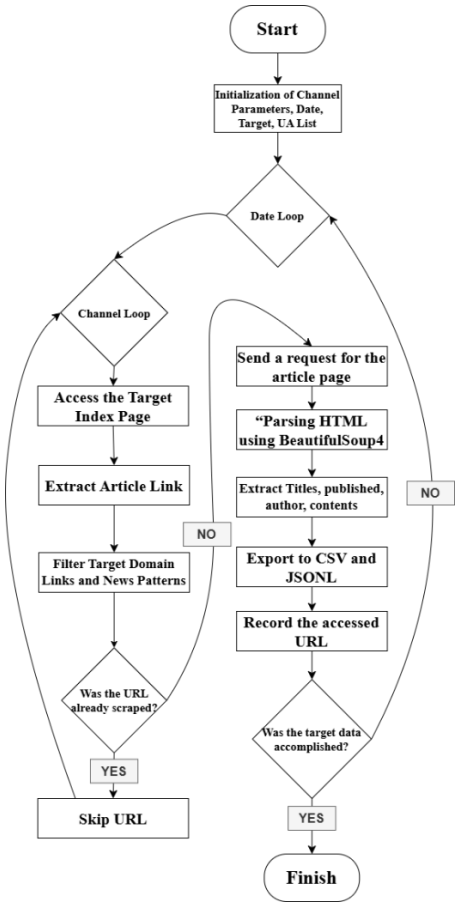

Figure 1 Hybrid lexical-semantic pipeline based on TF-IDF and indoBERT

The process begins with the extraction of news headlines and content exclusively from Detik.com using automated scraping techniques [9]. Each news entry consists of four primary attributes: title headline, publication date, author, and news content [9]. The acquired dataset undergoes a series of cleaning and tokenization stages to ensure a uniform text format, which is essential for improving the efficiency and accuracy of subsequent machine learning tasks [8], [10].

The subsequent stage involves the parallel extraction of lexical and semantic features [5]. Lexical features are obtained through Term Frequency-Inverse Document

Frequency (TF-IDF) analysis and token overlap counting, while semantic features are derived using IndoBERT embedding cosine similarity to capture deep contextual meanings [4], [6]. These representations are combined at the feature integration stage to form a hybrid lexical-semantic representation, which is then utilized as the input for a Single Layer Perceptron (SLP) model to determine the final clickbait classification [1].

B. Data Collection

The dataset used in this study was obtained through web scraping exclusively from Detik.com, one of the largest Indonesian online news portals. Data extraction was conducted using a structured approach to HTML elements such as <article>, <h1>, and <p> to ensure that the collected information accurately reflected the structural components of news articles [9]. The implementation of tools like BeautifulSoup allowed for the automated extraction of main textual content and associated metadata, which were subsequently stored in a structured format [2], [9]. The collected raw data were organized into columns including URL, Title, Author, Published Date, and Content [9].

A total of 3,000 news articles were collected, spanning political, entertainment, economic, and technological categories. To ensure systematic and reproducible labeling, clickbait and non-clickbait categories were assigned using a fully automated rule-based strategy. This labeling process was based on measurable lexical and semantic discrepancies between headlines and article content, as such misalignments are core characteristics of clickbait [1], [6]. Specifically, cosine similarity (TF-IDF), token overlap ratio, and IndoBERT cosine similarity were computed for each title-content pair to identify instances where the headline failed to reflect the article's actual information [2], [4].

Articles with similarity values below a predefined threshold were labeled as clickbait, while the remaining articles were labeled as non-clickbait. This approach follows a weak supervision paradigm, utilizing deterministic rules to generate labels based on observable semantic misalignment [2]. By relying on quantitative similarity thresholds, the labeling process ensures consistency and eliminates the subjective variability often found in manual inter-annotator processes [3], [5].

The final curated dataset consists of 3,000 labeled articles, comprising 432 clickbait and 2,568 non-clickbait instances, resulting in a class distribution of 14.40% and 85.60%, respectively. This indicates a class imbalance within the dataset, which was taken into consideration during model evaluation by emphasizing the F1-score alongside accuracy, precision, and recall metrics to provide a balanced measure of performance [3], [10]. Prior to the final labeling stage, deduplication and initial text filtering were performed to remove repeated articles from the same source to ensure data quality [4]. This procedure follows standard data cleaning practices commonly applied in clickbait detection research to improve model efficiency and reliability [2], [10]. Each filtered article was stored as a structured title-content pair to preserve semantic alignment throughout preprocessing and feature extraction stages [1], [6].

C. Stages of Text Pre-Processing

Text preprocessing was applied independently to both the title and content fields while maintaining document-level pairing. This stage aimed to standardize textual inputs, reduce noise, and ensure consistency prior to lexical and semantic feature extraction [8], [11].

1) *Data Cleaning and Text Normalization*: The initial cleaning stage involved converting all text into lowercase, a process known as case folding, to reduce token variation caused by capitalization differences [8], [11]. Lowercasing helps reduce data dimensionality while maintaining essential semantic information [8]. HTML tags, URLs, and multimedia markers were removed using pattern-based filtering to enhance the focus of the text [6], [10].

   Additionally, metadata prefixes commonly found in Indonesian online news articles, such as location and date indicators (e.g., “Jakarta -”, “Bandung, Senin (12/10/2025)”), were removed to prevent bias in similarity computation. Non-breaking space characters and excessive spacing were normalized to ensure consistent formatting, while unnecessary white space was removed to enhance dataset quality [12].
2) *Tokenization*: Tokenization was performed to separate the text into its smallest meaningful units, known as tokens [8]. This procedure was applied separately to both the cleaned title and content fields, ensuring that digits and punctuation marks were excluded to refine the input for subsequent analysis [8], [11]. This method allows the model to better understand the structural and contextual boundaries of the Indonesian news corpus [5].
3) *Stopword Removal*: *Stopword* removal was conducted using the StopWordRemoverFactory from the Sastrawi library, which is an NLP library specifically designed for the Indonesian language [8], [11]. This step eliminated frequently occurring function words that carry limited semantic weight-such as "yang," "dan," and "itu"-thereby reducing dimensionality, accelerating processing time, and decreasing noise in the analysis [8]. By focusing on keywords, this process improves the performance and accuracy of similarity calculations [5], [8].
4) *Stemming*: Stemming was applied using the Sastrawi library to normalize inflected and derived Indonesian words into their respective base or root forms [13]. This procedure involves the removal of affixes to reduce morphological variation, ensuring that semantically equivalent terms are mapped into consistent lexical representations [5], [13]. After stemming, empty tokens and single-character non-numeric tokens were removed to minimize residual noise and improve feature stability.
5) *Pair-Based Alignment Strategy*: Since clickbait in this study is operationally defined as a semantic discrepancy

between a headline and the article body, preprocessing was conducted on paired title-content representations [1], [2]. Clickbait headlines are often designed to exploit the curiosity gap by withholding information or providing hyperbolic phrases that do not fully reflect the internal content [4], [14]. Each news article was treated as a structured title-content pair to preserve the semantic alignment necessary for identifying such discrepancies [7]. This pairing structure was maintained throughout the pipeline to ensure that subsequent computations of lexical similarity (via TF-IDF and token overlap) and semantic similarity (via transformer-based embeddings) accurately reflected the alignment between headlines and their corresponding news content [3], [6].

D. Lexical and Semantic Feature Extraction

In this study, the lexical-semantic approach refers to a feature-level integration strategy that combines surface lexical similarity measures with contextual semantic representations [5], [6]. Lexical similarity captures explicit word-level alignment through statistical term weighting, while semantic similarity captures contextual meaning through the use of deep word or sentence embeddings [1], [2]. This integration is performed at the feature level to form a hybrid representation for downstream classification modeling, allowing the system to benefit from both frequency-based patterns and deep contextual relationships [5], [6].

1) Lexical Features: Lexical representation results from a combination of two main techniques:

- TF-IDF (Term Frequency-Inverse Document Frequency) was used to measure the relative importance of words within the news corpus [3], [6]. This method converts textual information into numerical vector forms suitable for machine learning inputs by weighing terms according to their frequency in a specific document relative to their rarity across the entire dataset [2], [10]. To ensure consistency in representation, a single TF-IDF vocabulary was fitted on the combined corpus of all titles and contents. This approach ensures that both fields are represented within the same feature space. After vectorization, the relationship between the headline and content vectors was evaluated. Lexical similarity was subsequently computed using cosine similarity to measure the degree of alignment based on word frequency distributions [8], [15].

$$\text{Cosine}(T, C) = \frac{T \cdot C}{\| T \| \| C \|}$$

Lower similarity values indicate a greater lexical discrepancy between the headline and the article body, which is a key indicator of clickbait behavior where the title fails to reflect the internal content [1], [2]. The TF-IDF computation was utilized to measure the significance of terms within the news corpus, converting textual information into numerical vector forms suitable for classification [6], [10]. TF-IDF computation was implemented using the Scikit-learn library with sublinear term frequency scaling and n-gram range (1-3) to capture both unigram and phrase-level patterns.

- The token overlap ratio was calculated by comparing the set of unique tokens in the headline with those in the corresponding news content. The overlap ratio was defined as:

$$\text{Overlap}(T, C) = \frac{| T \cap C |}{| T |}$$

where T and C denote the sets of unique stemmed tokens extracted from the title and content, respectively. The token overlap ratio was calculated by comparing the set of unique tokens in the headline with those in the corresponding news content. This metric provides a direct surface-level similarity measure independent of statistical weighting [5]. This approach follows the principle of n-gram matching to detect semantic alignment between news components (Jácobo-Morales & Marino-Jiménez, 2024). Together, these lexical features quantify the degree of explicit alignment between each headline and its corresponding article body [3], [6].

2) Semantic Features: For semantic representation, the pretrained indobenchmark/indobert-base-p2 model was used to generate contextual sentence embeddings. The model was applied as a static feature extractor without full fine-tuning to maintain computational efficiency and reduce model complexity, which is particularly beneficial when dealing with limited datasets or resource-constrained environments [1], [5]. Semantic similarity was computed using cosine similarity between the embedding vectors:

$$\text{Cosine}(E_T, E_C) = \frac{E_T \cdot E_C}{\| E_T \| \| E_C \|}$$

Where E_T and E_C represent the embedding vectors of the title and content, respectively. This score captures contextual alignment even when synonymous or paraphrased expressions are used, overcoming the limitations of frequency-based lexical methods [1], [5].

The lexical and semantic similarity scores were integrated at the feature level through direct concatenation with additional handcrafted features to form a unified hybrid feature vector [2], [6]. This concatenation strategy allows the model to leverage both surface-level lexical transparency and deep contextual semantic modeling [16]. The hybrid vector included additional numerical features such as word count, punctuation count (e.g., exclamation and question marks), and common clickbait or sensational keywords, which have been shown to have significant occurrence rate differences between clickbait and non-clickbait headlines [3], [14].

Prior to classification, the combined feature space was standardized using z-score normalization to ensure consistent

scaling across heterogeneous feature types and to prevent dominance of features with larger numeric ranges. This feature-level integration enables the simultaneous modeling of surface lexical alignment and contextual semantic discrepancy. The complete description of the extracted features is summarized in Table I.

TABLE I
DESCRIPTION OF FEATURES USED IN HYBRID LEXICAL-SEMANTIC APPROACH

Feature Type	Sub-Features	Description	Type	Source
Lexical	TF-IDF Cosine Similarity Weight	Cosine similarity between TF-IDF vectors of title and content	Numerical	TF-IDF Vectorization
	Token Overlap Ratio	Number of overlapping words	Numerical	Token Set Intersection
Semantic	IndoBERT Cosine Similarity	Cosine similarity between title and content embeddings	Numerical	IndoBERT Embedding
Extras	Title Length	Number of title tokens	Numerical	Token Count
	Punctuation Ratio	Ratio of punctuation symbols	Numerical	Regex Pattern
	Sensational Word Count	Count of lexicon-based sensational terms	Numerical	Lexicon Dictionary

E. Models and Architecture

The final stage of the pipeline involves classification using a Single Layer Perceptron (SLP), a linear classifier that operates without hidden layers. The concatenated hybrid feature vector serves as the input for a single linear transformation, followed by a sigmoid activation function to generate probability scores for binary clickbait classification [1]. The model was trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy as the loss function. Training was conducted with a batch size of 32 for 30 epochs. The SLP architecture was intentionally selected to evaluate the discriminative capability of the engineered hybrid features without the representational depth of multi-layer networks, thereby maintaining computational efficiency [5], [7].

In addition to the SLP, Logistic Regression (LR) and Random Forest (RF) were employed as non-neural comparator models. These algorithms are widely utilized in clickbait research as benchmarks due to their interpretability and effectiveness in high-dimensional feature spaces [7], [14]. All models were evaluated using accuracy, precision, recall, and F1-score metrics, which are standard performance measures for assessing the reliability of clickbait classification systems (Al-Sarem et al., 2021; Bronakowski et al., 2023).

F. Experimental Evaluation

Evaluation was conducted under five experimental scenarios to measure the contribution of different feature combinations to model performance:

1) TF-IDF + Token Overlap (Lexical Only)
2) TF-IDF + Token Overlap + Additional Features
3) BERT Only
4) BERT + Additional Features
5) TF-IDF + Token Overlap + BERT + Additional Features

To ensure robust evaluation and assess generalization performance, stratified 5-fold cross-validation was employed [1], [4]. This method partitions the dataset into five folds while preserving the original class distribution in each fold, preventing bias in the evaluation of imbalanced datasets [1], [3]. The dataset comprises 3,000 labeled articles, with 432 clickbait (14.40%) and 2,568 non-clickbait (85.60%) instances. Such class imbalances are common in Indonesian news corpora and require careful metric selection to avoid misleading accuracy scores [2], [3].

In each iteration, four folds were used for training and one-fold for validation, allowing for the generation of stable and unbiased performance estimates [1]. Aggregated confusion matrices were utilized to analyze classification consistency and identify common misclassification patterns, such as false positives driven by sensational terminology [5], [14]. While the dataset exhibits imbalance, no oversampling or undersampling techniques were applied in this specific workflow to maintain the natural data distribution of online news portals [2]. Instead, the F1-score was emphasized as the primary evaluation metric because it provides a harmonic mean of precision and recall, ensuring a balanced performance assessment in imbalanced classification settings [2], [3].

III. RESULTS AND DISCUSSION

A. Experimental Results

The experiment was conducted to evaluate the effect of combining lexical and semantic-based text preprocessing techniques on clickbait classification performance. Three classification algorithms were employed: Single Layer Perceptron (SLP), Logistic Regression (LR), and Random Forest (RF). Each model was evaluated using the following feature configurations:

1) TF-IDF+Token Overlap (lexical-only)

2) TF-IDF+Token Overlap+Additional Features
3) IndoBERT (semantic-only)
4) IndoBERT+Extra Features
5) TF-IDF+IndoBERT+Extra Features (hybrid lexical-semancic)

The dataset consists of 3,000 Indonesian news articles collected through web scraping from Detik.com portals. The dataset comprises 432 clickbait articles (14.40%) and 2,568 non-clickbait articles (85.60%), indicating a substantial class imbalance.

Model evaluation was conducted using stratified 5-fold cross-validation to ensure robust generalization assessment. In each fold, the original class distribution was preserved, and performance metrics were averaged across folds to obtain stable and unbiased estimates. The reported results represent the mean performance across all validation folds.

The preprocessing pipeline includes text cleaning, tokenization, stopword removal, and stemming using Sastrawi. Additional handcrafted features include title length, punctuation ratio, and sensational word count, based on prior findings that punctuation patterns and emotionally charged terms are commonly associated with clickbait.

Before model training, an exploratory analysis of similarity distributions between headlines and article content was conducted to examine semantic alignment differences across classes. The distribution of cosine similarity values is visualized in Figure 2.

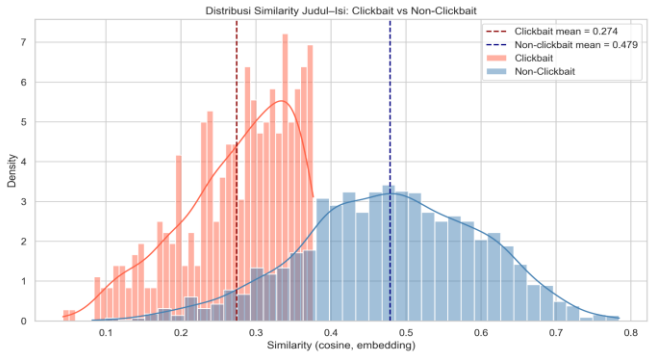

Figure 2 Visualization of cosine similarity value distribution

B. *Model Evaluation Results*

Table 4 shows the results of the model test based on accuracy, precision and recall metrics calculated using macro-averages to ensure a balance of assessments between classes (clickbait and non-clickbait).

TABLE II
CLICKBAIT CLASSIFICATION MODEL EVALUATION RESULTS

No	Feature Combinations	Models	*Accuracy*	*Precission*	*Recall*	*F1-Score*
1	TF-IDF + Token Overlap	SLP	0.86	0.87	0.85	0.86
2	TF-IDF + Token Overlap + Additional Features	SLP	0.88	0.89	0.87	0.88
3	BERT Only	SLP	0.83	0.82	0.85	0.83
4	BERT + Additional Features	SLP	0.85	0.86	0.84	0.85
5	TF-IDF + Token Overlap + BERT + Additional Features	SLP	0.87	0.88	0.86	0.87
6	TF-IDF + Token Overlap	LR	0.89	0.90	0.88	0.89
7	TF-IDF + Token Overlap + Additional Features	LR	0.90	0.91	0.90	0.90
8	BERT Only	LR	0.83	0.82	0.85	0.83
9	BERT + Additional Features	LR	0.84	0.85	0.83	0.84
10	TF-IDF + Token Overlap + BERT + Additional Features	LR	0.87	0.88	0.86	0.87
11	TF-IDF + Token Overlap	RF	0.98	0.98	0.98	0.98
12	TF-IDF + Token Overlap + Additional Features	RF	0.99	0.99	0.99	0.99
13	BERT Only	RF	0.98	0.98	0.98	0.98
14	BERT + Additional Features	RF	0.98	0.98	0.98	0.98
15	TF-IDF + Token Overlap + BERT + Additional Features	RF	1.00	1.00	1.00	1.00

Once all models are trained on a combination of different features, the results of precision metrics are tested to assess the model's ability to correctly predict clickbait. The results of the comparison of precision values can be seen in the following figure 3:

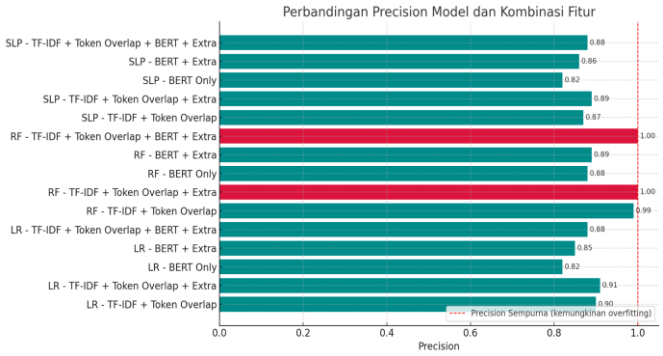

Figure 3 Precision Value Comparison Results

In addition to precision, recall metrics are also measured to find out the extent to which the model is able to recognize all clickbait instances in the test data. The results of the comparison of recall values between models and features are shown in the following figure 4:

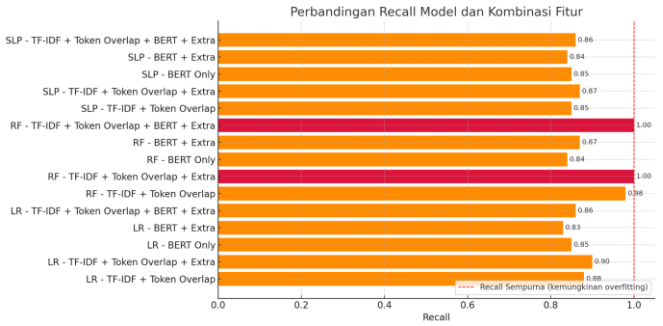

Figure 4 Results of comparison of recall values between models and features

C. *Model Performance Analysis*

Across the evaluated configurations, the SLP model demonstrates consistent and stable performance. The lexical-only configuration (TF-IDF + Token Overlap) achieved an F1-score of 0.86, indicating that lexical similarity effectively captures explicit alignment between headlines and article content.The addition of handcrafted features slightly improved performance to an F1-score of 0.88, suggesting that structural indicators such as title length and punctuation provide complementary information in detecting clickbait patterns.

The BERT-only configuration resulted in comparatively lower performance (F1-score of 0.83), indicating that semantic representation alone is insufficient without explicit lexical similarity signals. Adding additional handcrafted features to the BERT configuration did not yield substantial improvement, reinforcing the importance of combining semantic embeddings with lexical similarity rather than relying on contextual representation alone.

The hybrid configuration (TF-IDF + IndoBERT + Additional Features) achieved a balanced performance with an F1-score of 0.87 for SLP. These findings support the dual-representation paradigm widely adopted in modern NLP systems, where lexical and semantic features are integrated to enhance contextual understanding and reduce misclassification.

To further evaluate the trade-off between precision and recall, the F1-score metric was used as the primary indicator of balanced performance. A comparative visualization of F1-score values across models and feature combinations is presented in Figure 5.

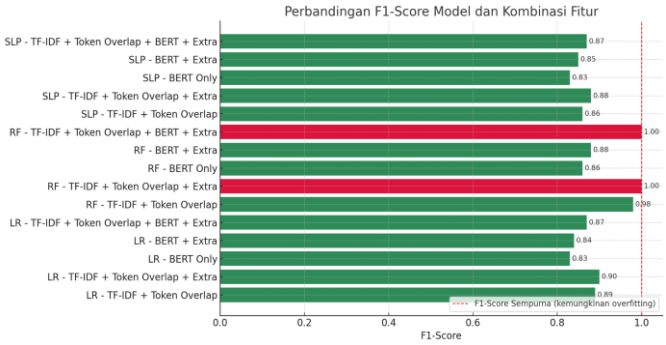

Figure 5 Comparison of F1-score values on each model and feature combination

D. *Comparative Analysis between Models (SLP, LR, RF)*

The hybrid lexical-semantic configuration was evaluated using three classification algorithms: Single Layer Perceptron (SLP), Logistic Regression (LR), and Random Forest (RF). From a numerical standpoint, Random Forest achieved the highest performance across multiple feature combinations, including a perfect F1-score (1.00) in the hybrid configuration.

However, this near-perfect performance should be interpreted carefully. Since the dataset labeling was generated through deterministic similarity-based weak supervision rules, the exceptionally high RF results may reflect the model's strong capacity to internalize structured feature-label relationships derived from those rules. Ensemble tree-based models are particularly effective at fitting nonlinear decision boundaries, which can lead to near-exact reconstruction of rule-based labeling patterns when feature-label alignment is highly structured.

In contrast, Logistic Regression demonstrated stable and consistent performance across configurations, particularly in lexical-based settings (F1-score up to 0.90). This suggests that linear decision boundaries are sufficiently expressive when similarity-driven features dominate the classification signal. The SLP model also achieved competitive and balanced results across configurations (F1-score up to 0.87), providing a lightweight neural alternative with comparable precision-recall trade-offs.

Overall, these findings indicate that performance variation across classifiers is influenced not only by model capacity but also by the structural dependency between engineered similarity features and the labeling mechanism. The observed improvements are therefore primarily attributable to feature integration rather than classifier complexity alone.

To further examine performance stability under class imbalance, macro and weighted F1-scores were compared for each model. The comparison results are presented in Figure 6.

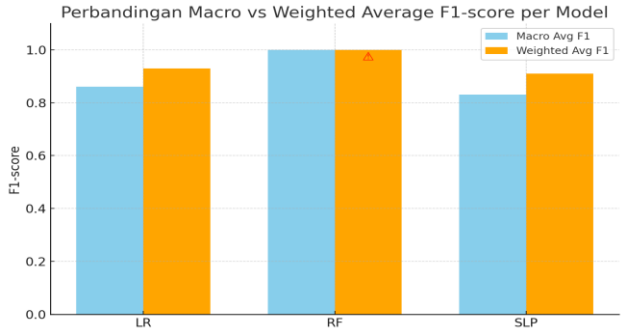

Figure 6 Comparison results between macro and weighted averager F1-score

Further evaluation was conducted on macro and weighted recall values to analyze sensitivity balance between clickbait and non-clickbait classes. The comparison is shown in Figure 7.

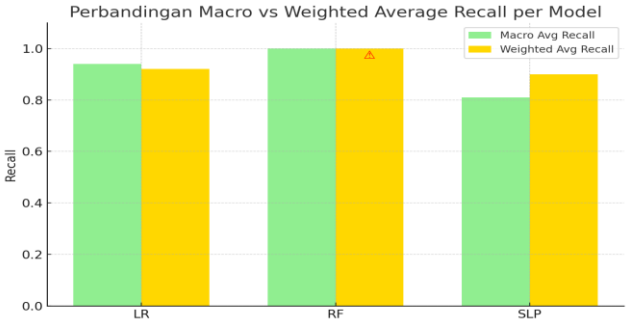

Figure 7 Comparison of macro and weighted recall values

E. Effectiveness of Hybrid Lexical-Semantic Approach

The proposed hybrid pipeline consistently improves classification robustness compared to single-representation approaches. While lexical-only configurations already provide strong performance, the integration of semantic embeddings further stabilizes performance across folds, particularly in cases where lexical overlap alone is insufficient to capture contextual alignment.

The hybrid configuration demonstrates a balanced trade-off between precision and recall, as reflected in its F1-score performance. This confirms that combining surface-level lexical similarity with contextual semantic representation enhances the model's ability to detect subtle headline-content discrepancies.

These findings support prior research indicating that dual representation strategies-integrating statistical and contextual features-improve robustness in text classification tasks. Importantly, this hybridization occurs at the feature-integration stage rather than relying solely on deeper neural architectures, offering a computationally efficient alternative suitable for practical deployment in online news environments.

F. Summary and Conceptual Discussion

Conceptually, the proposed Hybrid Lexical-Semantic Approach extends the Indonesian NLP processing paradigm by integrating TF-IDF statistical representations with IndoBERT contextual embeddings within a unified feature-level framework. This integration enables the system to simultaneously capture surface lexical similarity and deeper semantic alignment between headlines and content.

Rather than treating lexical and contextual models as competing approaches, this study demonstrates their complementary roles in clickbait detection. The evaluation framework, based on stratified cross-validation and multi-scenario feature configurations, strengthens methodological transparency and provides a stable empirical foundation for the proposed pipeline.

The consistent performance improvements across classifiers indicate that hybrid feature engineering contributes more significantly to classification effectiveness than classifier architecture alone.

G. Explainability and Feature Contribution Analysis

To address model interpretability, feature contribution analysis was conducted using the Random Forest classifier. The feature importance scores, as presented in Figure 8, indicate that semantic similarity (IndoBERT cosine similarity) is the most influential predictor, followed by lexical cosine similarity and token overlap.

As shown in Figure 8, semantic similarity contributes the largest proportion to the overall decision process, confirming that contextual misalignment between headline and article content is the primary indicator of clickbait in Indonesian news. While lexical overlap captures surface-level consistency, semantic similarity plays a dominant role in identifying subtle discrepancies where headlines are contextually misleading despite sharing vocabulary with the article body.

Additional handcrafted features, including sensational word count, title length, and punctuation ratio, contribute comparatively less to the classification outcome. This suggests that stylistic exaggeration alone is insufficient for reliable clickbait detection without modeling semantic alignment.

Furthermore, the extracted sensational lexicon derived from external labeled data provides linguistic interpretability. Terms with high log-odds ratios reveal statistically dominant clickbait patterns, often reflecting emotionally provocative or curiosity-inducing language constructions. These lexical patterns complement semantic modeling by highlighting rhetorical strategies commonly used in clickbait headlines.

Overall, the explainability analysis demonstrates that the proposed hybrid approach does not operate as a black-box model but instead integrates interpretable similarity-based signals and statistically validated lexical indicators, as empirically supported by the feature contribution visualization in Figure 8.

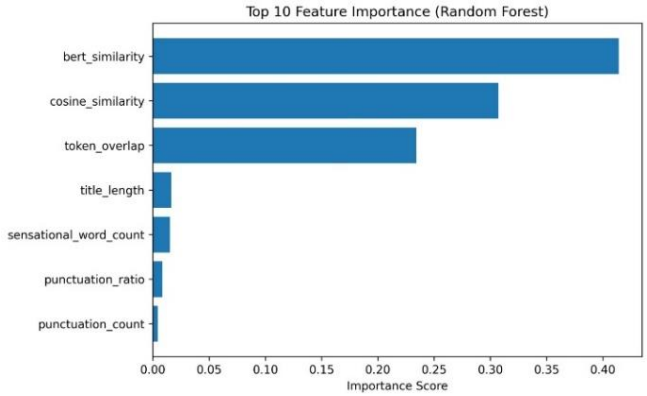

Figure 8 Feature Importance

H. Error Analysis

To further understand model behavior beyond aggregate metrics, qualitative error analysis was conducted on misclassified instances generated during cross-validation. Two primary types of errors were observed. The first category involves false positives, where non-clickbait headlines were predicted as clickbait. These cases typically exhibit low lexical overlap with the article body despite maintaining contextual relevance. For example, headlines that employ abstract or summarizing phrasing may share limited surface tokens with the content while remaining semantically aligned. In such cases, similarity-based features may underestimate alignment due to lexical compression.

The second category involves false negatives, where clickbait headlines were predicted as non-clickbait. These cases often contain emotionally suggestive or curiosity-driven wording while still maintaining moderate semantic similarity with the article content. Because the labeling strategy relies on measurable semantic discrepancy, rhetorically exaggerated headlines that remain contextually consistent may not be strongly penalized by similarity-based features.

Additionally, borderline instances were identified in which similarity values were close to the predefined quantile threshold. These cases highlight the inherent ambiguity in defining clickbait solely through lexical and semantic discrepancy, particularly when pragmatic or stylistic exaggeration is subtle.

Overall, the error analysis indicates that while the hybrid lexical-semantic approach effectively captures measurable headline-content misalignment, nuanced rhetorical clickbait patterns that rely on tone or discourse framing remain challenging under similarity-based supervision

I. Domain Limitation and Generalization Considerations

It is important to note that the dataset used in this study was collected exclusively from a single Indonesian news portal (Detik.com). While this ensures structural consistency and controlled annotation quality, it may introduce domain-specific bias associated with editorial style, headline construction patterns, and content formatting conventions.

Although stratified 5-fold cross-validation was applied to assess internal generalization performance, this evaluation strategy does not eliminate the possibility of cross-domain distribution shifts. Consequently, the trained models may partially capture stylistic characteristics unique to the source platform rather than universally generalizable clickbait patterns.

In addition, since labeling was conducted using automated similarity-based rules under a weak supervision paradigm, subtle rhetorical or contextual clickbait cases that do not exhibit measurable semantic discrepancy may not be fully captured. This may introduce a degree of labeling noise, particularly in headlines that rely on pragmatic ambiguity rather than lexical or semantic misalignment. Furthermore, exceptionally high classification scores observed in certain model configurations may partially reflect structural alignment between engineered similarity features and the rule-based labeling mechanism, rather than fully independent predictive generalization.

Therefore, the reported performance should be interpreted as domain-specific effectiveness within the Detik.com news environment and under similarity-based labeling assumptions. Future research should extend the proposed hybrid pipeline to multi-source and cross-domain datasets, as well as incorporate manual or hybrid annotation validation, to further assess robustness and external generalization capability across diverse Indonesian news ecosystems.

IV. CONCLUSION

This study demonstrates that a hybrid lexical-semantic framework integrating TF-IDF-based lexical similarity and IndoBERT-based semantic similarity provides a structured and computationally efficient approach for clickbait detection in Indonesian news. The empirical results indicate that feature-level integration contributes meaningfully to classification performance across different modeling configurations.

From a numerical perspective, the highest F1-score (0.90) was achieved in the lexical-based configuration using Logistic Regression, while the hybrid configuration produced stable and competitive performance (F1-score up to 0.87) across linear and lightweight neural classifiers. Although Random Forest achieved exceptionally high scores under certain configurations, these results should be interpreted cautiously given the deterministic similarity-based weak supervision used in the labeling process, which may increase structural alignment between engineered features and generated labels.

Overall, the findings suggest that performance gains are primarily driven by representation design rather than classifier complexity alone. Lexical similarity features effectively capture explicit headline-content alignment, whereas contextual semantic embeddings contribute to modeling deeper alignment patterns. The feature-level concatenation strategy enables complementary interaction between surface statistical signals and contextual representations within a computationally feasible architecture.

The evaluation was conducted under a stratified 5-fold cross-validation framework to ensure stable internal performance estimation. Distinct feature subsets were

systematically examined across experimental configurations to maintain transparency in feature-target relationships. Nevertheless, several limitations remain. The dataset was collected exclusively from a single Indonesian news portal, which may introduce domain-specific stylistic bias. In addition, the weak supervision labeling strategy-based on similarity thresholds-may not fully capture pragmatic or discourse-level clickbait phenomena that extend beyond measurable lexical or semantic discrepancy.

Future research should evaluate the proposed framework on multi-source, cross-domain, and multi-country news datasets to assess broader external generalization beyond a single national platform. Incorporating manually validated annotations from domain experts or professional news evaluators would improve label reliability and reduce dependency on rule-based weak supervision. In addition, applying appropriate label balancing strategies-such as resampling techniques or cost-sensitive learning-may enhance stability when dealing with substantial class imbalance. Exploring discourse-level or pragmatic features could further strengthen the detection of nuanced clickbait patterns that extend beyond measurable lexical or semantic discrepancy.

REFERENCES

[1] M. N. Fakhruzzaman, S. Z. Jannah, R. A. Ningrum, and I. Fahmiyah, "Clickbait Headline Detection in Indonesian News Sites using Multilingual Bidirectional Encoder Representations from Transformers (M-BERT)," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.01497

[2] M. Al-Sarem *et al.*, "An improved multiple features and machine learning-based approach for detecting clickbait news on social networks," *Applied Sciences (Switzerland)*, vol. 11, no. 20, Oct. 2021, doi: 10.3390/app11209487.

[3] A. Muqadas, H. U. Khan, M. Ramzan, A. Naz, T. Alsahfi, and A. Daud, "Deep learning and sentence embeddings for detection of clickbait news from online content," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-97576-1.

[4] J. Sirusstara, N. Alexander, A. Alfarisy, S. Achmad, and R. Sutoyo, "Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa)," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences: Championing Innovations in Artificial Intelligence and Data Sciences for Sustainable Future, AiDAS 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 248-253. doi: 10.1109/AiDAS56890.2022.9918678.

[5] Sutriawan, S. Rustad, G. F. Shidik, and Pujiono, "Performance Evaluation of Text Embedding Models for Ambiguity Classification in Indonesian News Corpus: A Comparative Study of TF-IDF, Word2Vec, FastText BERT, and GPT," *Ingenierie des Systemes d'Information*, vol. 30, no. 6, pp. 1469-1482, Jun. 2025, doi: 10.18280/isi.300606.

[6] W. Du, C. Ge, S. Yao, N. Chen, and L. Xu, "Applicability Analysis and Ensemble Application of BERT with TF-IDF, TextRank, MMR, and LDA for Topic Classification Based on Flood-Related VGI," *ISPRS Int. J. Geoinf.*, vol. 12, no. 6, Jun. 2023, doi: 10.3390/ijgi12060240.

[7] A. Chowanda, Nadia, and L. M. M. Kolbe, "Identifying clickbait in online news using deep learning," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1755-1761, Jun. 2023, doi: 10.11591/eei.v12i3.4444.

[8] D. Iskandar and A. Kurniawati, "Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 133-144, Feb. 2025, doi: 10.25126/jtiik.2025129143.

[9] A. Khanom, D. Kiesow, M. Zdun, and C. R. Shyu, "The News Crawler: A Big Data Approach to Local Information Ecosystems," *Media Commun.*, vol. 11, no. 3, pp. 318-329, 2023, doi: 10.17645/mac.v11i3.6789.

[10] N. Sardana, D. Varshney, and S. Luthra, "Enhanced Clickbait Detection through Ensemble Machine Learning Techniques," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 599-608. doi: 10.1016/j.procs.2025.04.294.

[11] 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, Aug. 2024, doi: 10.25126/jtiik.1148315.

[12] M. A. Taha, H. D. A. Jabar, and W. K. Mohammed, "A Machine Learning Algorithms for Detecting Phishing Websites: A Comparative Study," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 3, pp. 275-286, 2024, doi: 10.52866/ijcsm.2024.05.03.015.

[13] F. Rumaisa, "Evaluation of Indonesian Language Stemmer Algorithms: A Comparative Analysis," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 1, pp. 21-25, Mar. 2025, doi: 10.47709/brilliance.v5i1.5679.

[14] M. Bronakowski, M. Al-khassaweneh, and A. Al Bataineh, "Automatic Detection of Clickbait Headlines Using Semantic Analysis and Machine Learning Techniques," *Applied Sciences (Switzerland)*, vol. 13, no. 4, Feb. 2023, doi: 10.3390/app13042456.

[15] R. N. Tanaja, Johnny, M. A. Rafif, and A. A. S. Gunawan, "Fake News Detection using Machine Learning: Integrating FakeBERT Classification, Style Analysis, and Credibility Verification," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 1067-1076. doi: 10.1016/j.procs.2025.09.048.

[16] A. Hashemi, M. R. Moosavi, W. Shi, and A. Giachanou, "Enhancing fake news detection through estimating user tendencies to spread fake news," *Data Inf. Manag.*, vol. 10, no. 2, Jun. 2026, doi: 10.1016/j.dim.2025.100115.