International Journal Software Engineering and Computer Science (IJSECS)
6 (3), 2026, 1168-1184
Published Online December 2026 in IJSECS (http://www.journal.lembagakita.com/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v6i3.8156.

RESEARCH ARTICLE Open Access

Developing an Indonesian Fake News Detection System Using IndoBERT and ProtoNet for Few-Shot Learning

Yefta Christian [1*], **Wilson** [2], **Andik Yulianto** [3]

[1*,2,3] Universitas International Batam, Batam City, Riau Islands Province, Indonesia.

Article History:

Received: July 23, 2026.
Revised: July 25, 2026.
Accepted: August 27, 2026.
Available online: September 20, 2026.
Published: December 1, 2026.

***Corresponding Author:**
yefta@uib.ac.id

Abstract: Fake news dissemination through digital media has become a critical concern because inaccurate information may spread swiftly and influence public comprehension, social behavior, and decision-making. This study developed an Indonesian fake news detection web application employing fine-tuned IndoBERT and an IndoBERT-based Prototypical Network (ProtoNet) under few-shot learning scenarios. The research followed the Cross Industry Standard Process for Data Mining framework, consisting of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The dataset was obtained from the publicly available Deteksi Berita Hoaks Indo Dataset on Kaggle and contained 23,944 Indonesian news records, consisting of 11,200 factual and 12,744 hoax articles. Three few-shot scenarios were evaluated: 50-shot, 100-shot, and 300-shot per class. IndoBERT was fine-tuned as a supervised classification baseline, while ProtoNet used IndoBERT embeddings to construct class prototypes and classify texts based on Euclidean distance. The best performance was achieved by fine-tuned IndoBERT in the 300-shot scenario, with 97.12% accuracy and a 97.10% macro F1-score. ProtoNet achieved 92.65% accuracy and a 92.63% macro F1-score in the same scenario. Even in the most constrained 50-shot setting, ProtoNet achieved an F1-score of 90.82%, closely trailing fine-tuned IndoBERT at 91.17% and demonstrating strong metric stability. The web application supports manual text input, URL-based article extraction, model selection, confidence scores, and word-level occlusion explanations. These results demonstrate the effectiveness of IndoBERT for Indonesian fake news detection and the feasibility of prototype-based classification under limited-data conditions. Overall, this work bridges the gap between deep contextual representation, data-efficient learning, and practical, interpretable deployment for public media verification.

Keywords: Fake News Detection; IndoBERT; Prototypical Network; Few-Shot Learning; Explainable Artificial Intelligence.

1. Introduction

The rapid development of digital media has significantly changed how people access, consume, and distribute information. Online news portals, social media platforms, and messaging applications allow information to spread quickly across diverse user communities. While this environment provides immediate access to news, it simultaneously heightens the risk of fake news dissemination. Misleading articles intentionally mimic legitimate journalism to deceive readers, making content-based detection both difficult and non-trivial (Shu *et al.*, 2017). Furthermore, fake news detection extends beyond simple keyword matching. As highlighted by Zhou and Zafarani (2020), detection encompasses multiple dimensions, including false knowledge, writing styles, propagation patterns, and source credibility. Nevertheless, in practical web systems, textual content remains the most accessible input because articles can be directly ingested via manual submission or web extraction. Consequently, automated text-based detection remains an essential research direction. In the Indonesian context, automated fake news detection presents specific linguistic and operational challenges. Fawaid *et al.* (2021) observed that deceptive information spreads extensively in Indonesia despite limited computational resources for Bahasa Indonesia, demonstrating that Transformer-based architectures outperform conventional CNN, BiLSTM, and hybrid CNN-BiLSTM models. Similarly, Azizah *et al.* (2023) evaluated Transformer variants and confirmed that contextual models such as ALBERT and BERT deliver superior classification performance over traditional machine learning baselines. These findings indicate that Transformer-based language representations are well suited for identifying linguistic anomalies in Indonesian misinformation. Transformer-based models have established strong empirical benchmarks in natural language processing through multi-head self-attention mechanisms that eliminate recurrence and convolution (Vaswani *et al.*, 2017). Devlin *et al.* (2019) extended this capability through BERT, enabling bidirectional pretraining that can be readily fine-tuned for downstream classification tasks with minimal architectural modifications. For Indonesian text processing, language-specific models are essential for capturing local syntactic nuances and colloquial structures. Koto *et al.* (2020) demonstrated the effectiveness of IndoBERT across multiple benchmark tasks within the IndoLEM framework, while Wilie *et al.* (2020) established IndoNLU and released foundational models trained on the extensive Indo4B corpus. These resources support the adoption of IndoBERT as an effective backbone for Indonesian text classification. Although fully supervised fine-tuning yields high classification accuracy, real-world misinformation detection frequently encounters limited labeled data. Verifying news veracity requires rigorous fact-checking, trusted primary sources, and specialized domain knowledge. Under such constraints, few-shot learning offers a practical alternative by enabling models to generalize from a minimal number of labeled instances (Wang *et al.*, 2020). Among metric-based meta-learning frameworks, Prototypical Networks (ProtoNet) represent each class through a prototype computed as the mean embedding of its support examples, classifying query instances by measuring Euclidean distances to these class centroids (Snell *et al.*, 2017). In few-shot text classification, inductive representations and encoder quality strongly govern cross-task generalization and class separation (Bao *et al.*, 2019; Dopierre *et al.*, 2021; Geng *et al.*, 2019). Leveraging IndoBERT as a representation encoder alongside a metric-based prototype classifier thus provides a compelling framework for limited-data scenarios. Another critical requirement for deployed detection systems is interpretability. Complex deep learning models function largely as black boxes, creating operational friction when end users must evaluate model credibility (Arrieta *et al.*, 2020; Ribeiro *et al.*, 2016). While feature attribution techniques such as SHAP provide unified post-hoc importance metrics (Lundberg & Lee, 2017), the internal reasoning mechanisms of BERT-based models remain challenging to isolate completely (Rogers *et al.*, 2020). Therefore, explainability mechanisms in practical tools should serve as transparent interpretive aids rather than definitive proof of internal model reasoning. Despite these advancements, existing Indonesian fake news detection research has predominantly focused on fully supervised learning with large annotated corpora, leaving few-shot classification, web deployment, and user-oriented explainability largely unaddressed. To resolve these gaps, this study develops an Indonesian fake news detection web application integrating fine-tuned IndoBERT and an IndoBERT-based Prototypical Network under few-shot learning conditions. The novelty of this work lies in combining supervised fine-tuning, prototype-based metric learning, URL-based article parsing, and word-level occlusion explanations within a unified web interface, extending prior work on web-integrated Indonesian NLP models (Fakhruzzaman & Gunawan, 2021). Specifically, this study aims to: (1) develop a web-based Indonesian fake news detection application; (2) implement fine-tuned IndoBERT for Indonesian fake news classification; (3) implement a Prototypical Network using IndoBERT embeddings under few-shot scenarios; (4) evaluate and compare model performance under 50-shot, 100-shot, and 300-shot configurations; and (5) incorporate explainable artificial intelligence functionality using word-level occlusion.

2. Related Work

2.1 Fake News Detection

Fake news detection has been examined as a computational problem involving textual content, writing style, source credibility, and information propagation. Shu *et al*. (2017) explained that fake news is intentionally constructed to mislead readers, making it difficult to identify through simple keyword-based methods. Zhou and Zafarani (2020) subsequently categorized fake news detection perspectives into false knowledge, writing style, propagation patterns, and source credibility, indicating that detection requires deeper contextual and semantic signals rather than surface lexical markers. Content-based detection remains particularly practical for web deployment because article text can be ingested directly through user input or automated URL scraping. However, text classification remains challenging when deceptive articles deliberately mimic legitimate journalistic style and vocabulary. Recent Indonesian studies have highlighted the effectiveness of IndoBERT in handling local linguistic characteristics. Simanjuntak *et al*. (2024) evaluated Bayesian optimization, grid search, and random search for IndoBERT fine-tuning, demonstrating that Bayesian optimization yielded an F1-score of 91.56% for the hoax category. Addressing class imbalance in Indonesian political news, Fathin *et al*. (2024) applied random oversampling and undersampling, attaining 98.2% accuracy and a 97.8% F1-score; however, their framework remained strictly fully supervised without exploring low-resource conditions. Hybrid approaches have also shown promising results: Yefferson *et al*. (2024) paired IndoBERT embeddings with an LSTM classifier on 5,876 articles to achieve an F1-score of 90.8%, while Prisscilya and Girsang (2024) integrated BERTopic with IndoBERT to incorporate thematic features, achieving a test F1-score of 91%. Furthermore, Pekandi *et al*. (2025) benchmarked IndoBERT against ensemble and CNN-LSTM models across 6,120 articles, obtaining an F1-score of 92.23%, and Jocelynne *et al*. (2025) demonstrated competitive performance in political hoax classification. While these studies underscore the utility of language-specific Transformer models, they collectively focused on conventional supervised training on abundant data rather than metric-based few-shot learning.

2.2 Transformer, BERT, and IndoBERT

The Transformer architecture introduced by Vaswani *et al*. (2017) replaced recurrent sequence modeling with self-attention mechanisms, allowing models to capture non-local contextual dependencies efficiently in parallel. Devlin *et al*. (2019) subsequently developed BERT through bidirectional masked language modeling and next sentence prediction, establishing a paradigm where general semantic representations are pretrained on large-scale corpora and subsequently fine-tuned for downstream tasks with minimal architectural additions. Although multilingual BERT can process Indonesian text, dedicated monolingual models capture morphological richness and informal syntactic patterns more effectively. Koto *et al*. (2020) introduced IndoBERT alongside the IndoLEM benchmark, while Wilie *et al*. (2020) established IndoNLU and released foundational models trained on the expansive Indo4B corpus. Existing Indonesian fake news research confirms that IndoBERT yields high classification accuracy across diverse architectures. However, performance remains sensitive to data distribution, hyperparameter optimization, and downstream classification structures. In this research, IndoBERT serves a dual role: first as an end-to-end supervised baseline via classification head fine-tuning, and second as a contextual feature extractor generating embeddings for a Prototypical Network. This dual formulation establishes a rigorous empirical comparison between parametric supervised fine-tuning and non-parametric metric learning using an identical Indonesian linguistic backbone.

2.3 Few-Shot Learning and Prototypical Networks

Few-shot learning addresses scenarios where supervised labeled data are scarce, a condition frequently encountered in misinformation detection due to the substantial human labor required for factual verification (Wang *et al*., 2020). Snell *et al*. (2017) formulated Prototypical Networks (ProtoNet) as a metric-based meta-learning technique where each class is represented by a prototype vector calculated as the mean embedding of its support instances, and query points are assigned labels according to their Euclidean distance to these centroids. Subsequent investigations have refined prototype-based text classification under complex data distributions. Sehanobish *et al*. (2022) developed Variance-Aware Prototypical Networks to represent class prototypes as probabilistic distributions, improving out-of-distribution robustness across multiple text datasets. To address semantic overlap between closely related classes, Lei *et al*. (2023) proposed Task-Adaptive Reference Transformation (TART), showing substantial accuracy gains on news classification benchmarks. In addition, X. Liu *et al*. (2024) applied optimal transport to alleviate prototype bias stemming from high intra-class variance, whereas Wen *et al*. (2025) proposed GAProtoNet, integrating graph attention with Transformer embeddings to provide competitive classification accuracy alongside prototype interpretability. These developments emphasize that encoder quality and class separation dictate few-shot generalization, supporting the use of IndoBERT as the embedding foundation evaluated under 50-shot, 100-shot, and 300-shot configurations.

2.4 Explainable Artificial Intelligence

Explainable Artificial Intelligence provides mechanisms that elucidate machine learning predictions, mitigating black-box opacity in sensitive applications (Arrieta *et al.*, 2020). Ribeiro *et al.* (2016) established LIME to explain individual predictions using interpretable local surrogates, demonstrating that classification decisions require interpretable evidentiary support. In misinformation detection, Lu and Li (2020) utilized graph co-attention to highlight influential tokens and suspicious social accounts, while Han *et al.* (2024) developed a multifaceted reasoning network to capture granular contextual cues. Furthermore, H. Liu *et al.* (2024) established TELLER to derive structured reasoning rules alongside classification labels, and Wen *et al.* (2025) related predictions directly to representative prototypical exemplars. Despite these advancements, Transformer representations remain challenging to interpret fully (Rogers *et al.*, 2020). In this study, word-level occlusion is implemented as a direct, post-hoc explanation technique. By quantifying the degradation in prediction confidence when individual tokens are omitted, the system identifies the most influential terms driving the output. This approach conveys prediction sensitivity transparently without making unverified claims about internal model reasoning.

2.5 Web-Based System Architecture

Deploying machine learning models via web interfaces enables accessible end-user interaction without requiring local execution environments. Modular web architectures decouple client interaction, text extraction, preprocessing, model inference, and explanation components. For instance, Erlina *et al.* (2023) demonstrated the feasibility of deploying AI components into online service environments via automated web interfaces. Extending this principle to misinformation detection, the system developed in this study integrates fine-tuned IndoBERT and IndoBERT-based ProtoNet into a unified FastAPI backend. The application processes both direct text input and external news URLs via automated body extraction, returning prediction labels, confidence scores, and word-level occlusion highlights. This modular design permits dynamic model updating while allowing users to compare supervised and metric-based models through a single interactive interface. A comparative summary of related works is presented in Table 1.

2.6 Comparison with Previous Studies

To systematically position this research within the existing literature, previous studies in Indonesian fake news detection, few-shot text classification, and explainable misinformation detection are compared across key dimensions: method architecture, dataset characteristics, primary performance outcomes, methodological strengths, and operational limitations. As summarized in Table 1, while previous Indonesian studies have established the efficacy of IndoBERT under conventional supervised learning, they rely heavily on large-scale annotated datasets and do not examine performance under low-resource constraints. Conversely, recent metric-based few-shot frameworks have largely been validated on English or general domain-specific benchmarks rather than Indonesian news corpora. The proposed study bridges these domains by pairing fine-tuned IndoBERT with an IndoBERT-based Prototypical Network under rigorous few-shot scenarios, embedding the models within an explainable web platform supporting real-time URL article parsing.

Table 1. Comparison with Previous Studies

Study	Method	Dataset	Main Results	Strength	Limitation
Simanjuntak *et al.* (2024)	IndoBERT with hyperparameter optimization	Indonesian fake-news data	F1-score of 91.56% for the fake class	Examined several tuning approaches	No few-shot evaluation
Fathin *et al.* (2024)	IndoBERT with ROS and RUS	6,947 hoax and 20,945 factual articles	Accuracy 98.2%; F1-score 97.8%	Addressed class imbalance	Required a large supervised dataset
Yefferson *et al.* (2024)	IndoBERT-LSTM	5,876 Indonesian articles	Accuracy 93.2%; F1-score 90.8%	Combined contextual and sequential representations	No Prototypical Network
Prisscilya & Girsang (2024)	BERTopic and IndoBERT	Indonesian false-news dataset	Test F1-score 91%	Incorporated topic information	Not evaluated under limited-data settings
Pekandi *et al.* (2025)	IndoBERT, ensemble, and CNN-LSTM	6,120 Indonesian articles	IndoBERT F1-score 92.23%	Compared multiple model families	Conventional supervised learning

Jocelynne *et al.* (2025)	Fine-tuned IndoBERT	Indonesian political-news data	Evaluated using accuracy, precision, recall, F1, and ROC-AUC	Focused on Indonesian political hoaxes	No few-shot classification
Sehanobish *et al.* (2022)	Variance-Aware Prototypical Network	17 text datasets	Outperformed strong baselines	Modeled prototype uncertainty	Not Indonesian and not fake news
Lei *et al.* (2023)	TART few-shot classification	Four benchmark datasets	Improvements in one-shot and five-shot settings	Improved class separation	Not designed for Indonesian text
X. Liu *et al.* (2024)	Task-adaptive metric learning and optimal transport	Eight text datasets	Outperformed evaluated few-shot baselines	Improved prototype estimation	More complex than standard ProtoNet
Wen *et al.* (2025)	Transformer and graph-attention ProtoNet	Multiple text datasets	Competitive accuracy and F1-score	Combined prototypes and interpretability	Not evaluated for Indonesian fake news
H. Liu *et al.* (2024)	TELLER	Four fake-news datasets	Improved trustworthy and explainable detection	Generated human-readable reasoning components	Requires a more complex reasoning pipeline
Proposed study	Fine-tuned IndoBERT and IndoBERT-based ProtoNet	Indonesian fake and factual news (23,944 articles)	Evaluated in 50-, 100-, and 300-shot settings	Integrates supervised learning, few-shot learning, web deployment, URL extraction, and explanations	Explanations represent sensitivity rather than complete model reasoning

2.7 Research Positioning

Previous studies demonstrate that IndoBERT serves as an effective foundation for Indonesian fake news detection. However, prevailing research has almost exclusively relied on fully supervised learning with large annotated corpora. Conversely, while recent few-shot text classification research has introduced sophisticated metric-space adaptations, these techniques have primarily been evaluated on standard English benchmarks rather than Indonesian misinformation datasets. Furthermore, few studies have investigated Indonesian pretrained language models as metric encoders within Prototypical Networks. Prior literature has also treated model training, web deployment, URL-based content extraction, and post-hoc interpretability as disparate engineering challenges rather than integrating them into a cohesive verification system. This study directly addresses these gaps by evaluating IndoBERT under both parametric fine-tuning and metric-based Prototypical Network configurations across 50-shot, 100-shot, and 300-shot scenarios, embedding both paradigms into an accessible, explainable web application.

3. Methodology

3.1 Research Framework

This study adopts the Cross Industry Standard Process for Data Mining (CRISP-DM) framework to structure the development, empirical evaluation, and operational deployment of the detection system. Christian and Yap Rui Qi (2022) previously demonstrated the efficacy of CRISP-DM in guiding machine learning workflows across iterative project cycles. As illustrated in Figure 1, the research methodology encompasses six sequential phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The business understanding stage establishes the foundational objective of automating Indonesian fake news identification under low-resource constraints. Data understanding analyzes the structural characteristics, linguistic attributes, and class balance of the acquired news corpus. Data preparation performs text sanitization, column normalization, stratified partitioning, and balanced few-shot sampling. In the modeling stage, fine-tuned IndoBERT and prototype-based Prototypical Networks are parameterized across varying shot settings. The evaluation phase quantifies model performance via multi-class classification

metrics on an isolated test set. Finally, the deployment phase integrates the trained model weights into an interactive web application equipped with automated content scraping and post-hoc interpretability.

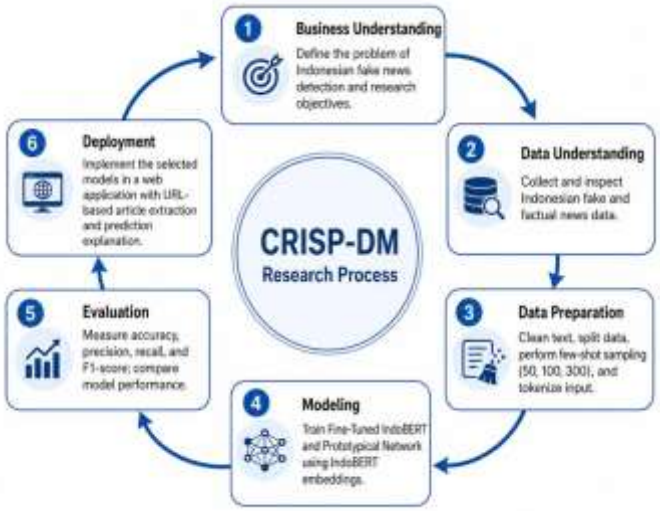

Figure 1. Research framework using CRISP-DM

3.2 Dataset Description

The experimental corpus was obtained from the publicly accessible Deteksi Berita Hoaks Indo Dataset hosted on Kaggle. The dataset comprises 23,944 verified Indonesian news articles, categorized into 11,200 factual reports (46.78%) and 12,744 hoax articles (53.22%), exhibiting a relatively balanced class distribution. The primary attributes utilized in the pipeline are clean_text, containing the sanitized article body, and label, denoting the binary ground-truth veracity class. To prevent data leakage while preserving class proportions, the dataset was partitioned into training, validation, and testing subsets using a stratified 70:10:20 split ratio. As summarized in Table 2, the training partition comprises 16,760 articles (7,840 factual and 8,920 hoax), the validation set contains 2,395 articles (1,120 factual and 1,275 hoax), and the held-out testing set includes 4,789 articles (2,240 factual and 2,549 hoax). The training partition served as the sampling pool for generating the balanced few-shot subsets, the validation set was utilized for epoch-level performance tracking and hyperparameter checkpointing, and the testing set was reserved exclusively for final model benchmarking.

Table 2. Dataset Distribution

Dataset Subset	Factual	Hoax	Total	Proportion
Training	7,840	8,920	16,760	70%
Validation	1,120	1,275	2,395	10%
Testing	2,240	2,549	4,789	20%
Total	11,200	12,744	23,944	100%

3.3 Data Preparation

The data preparation pipeline involved removing non-textual metadata attributes, filtering out missing values to ensure tokenization stability, standardizing column identifiers, and partitioning the corpus into stratified training, validation, and test subsets. For the few-shot learning experiments, balanced support subsets were sampled from the training partition using a fixed random seed of 42 to guarantee experimental reproducibility. Across both fine-tuned IndoBERT and Prototypical Network architectures, three distinct few-shot configurations were evaluated, as detailed in Table 3: 50 samples per class, 100 samples per class, and 300 samples per class. Input texts were tokenized using the pretrained IndoBERT tokenizer, with sequences padded and truncated to a uniform maximum sequence length of 128 tokens.

Table 3. Few-Shot Scenarios

Scenario	Samples per Class	Purpose
50-shot	50 samples per class	To evaluate model performance with very limited labeled data
100-shot	100 samples per class	To evaluate performance with moderate few-shot data
300-shot	300 samples per class	To evaluate performance with larger limited-data setting

3.4 Fine-Tuning IndoBERT

The first classification strategy employed parametric fine-tuning of the indobenchmark/indobert-base-p1 model, instantiated with a sequence classification head over two target classes. Fine-tuning was executed independently across the 50-shot, 100-shot, and 300-shot subsets. Each model configuration was trained for three epochs utilizing the AdamW optimizer with a learning rate of 2×10^{-5}, an effective batch size of 8, and cross-entropy loss as the optimization objective. Model checkpoints were evaluated and recorded at the conclusion of each training epoch, with the checkpoint achieving the highest validation macro F1-score subsequently restored for final benchmarking. The fine-tuning workflow proceeds sequentially: input tokens are processed by the IndoBERT encoder, after which the pooled representation is mapped through the linear classification head to output class probabilities.

3.5 Validation Strategy and Overfitting

Monitoring Generalization performance for the fine-tuned IndoBERT models was monitored across epochs using the dedicated validation partition of 2,395 articles. Loss values, accuracy, macro precision, macro recall, and macro F1-scores were tracked to evaluate convergence stability. Monitoring was particularly critical in the 50-shot scenario where minimal training instances increase susceptibility to empirical overfitting. As presented in Table 4, the training loss decreased steadily from 0.5030 at Epoch 1 to 0.0641 at Epoch 3. Concurrently, the validation loss declined from 0.3325 to 0.2593, while the validation macro F1-score improved from 0.8822 to 0.9165. This simultaneous reduction in validation error confirms stable parameter convergence without evidence of catastrophic overfitting.

Table 4. Training and Validation Performance in the 50-Shot IndoBERT Scenario

Epoch	Training Loss	Validation Loss	Validation Macro F1
1	0.5030	0.3325	0.8822
2	0.1687	0.2639	0.9137
3	0.0641	0.2593	0.9165

3.6 Prototypical Network with IndoBERT Embedding

The second classification strategy implemented a Prototypical Network utilizing the frozen indobenchmark/indobert-base-p1 encoder to construct dense metric representations. For each input article, the contextual embedding corresponding to the [CLS] token was extracted from the final hidden layer. Using balanced support sets drawn under the 50-shot, 100-shot, and 300-shot configurations, class prototype vectors were established by calculating the mean embedding across support instances within each respective veracity category. Query instances were subsequently classified by computing the Euclidean distance to each class prototype and assigning the label of the nearest prototype centroid, in accordance with the metric-learning framework of Snell *et al.* (2017). The ProtoNet encoder was optimized over three training epochs using the Adam optimizer with a learning rate of 2×10^{-5}, a batch size of 4, and gradient checkpointing enabled to mitigate memory overhead. The objective function minimized cross-entropy over the negative Euclidean distances. Unlike the fine-tuned IndoBERT baseline, the ProtoNet implementation was evaluated directly using the terminal model parameters on the testing partition without an intermediate validation checkpoint selection stage. The complete operational pipeline, illustrating both IndoBERT fine-tuning and Prototypical Network metric derivation, is depicted in Figure 2.

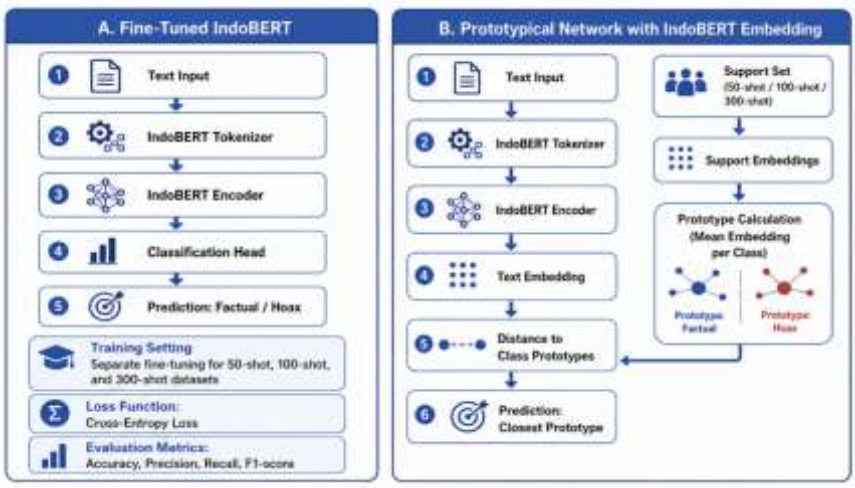

Figure 2. IndoBERT and Prototypical Network workflow

3.7 Web Application Design

The proposed detection system was operationalized as a modular, service-oriented web application comprising a presentation interface, a RESTful API backend, a model inference pipeline, an automated content parsing module, and an explainability engine. As demonstrated by Christian and Yap Rui Qi (2022), decoupling client-side interfaces from computational backends facilitates practical accessibility for non-technical users. The frontend provides input controls for manual text submission, external article URL ingestion, architecture selection (IndoBERT versus ProtoNet), and shot configuration selection (50-shot, 100-shot, or 300-shot). Upon receiving a request, the backend API coordinates the inference pipeline. When a URL is submitted, the newspaper3k parsing service extracts the article headline and body text, synthesizing them into a standardized textual payload. The backend subsequently passes the processed tokens through the selected model checkpoint, returning the predicted class label, associated confidence score, and word-level occlusion attributions. The overall structural architecture of the web platform is illustrated in Figure 3.

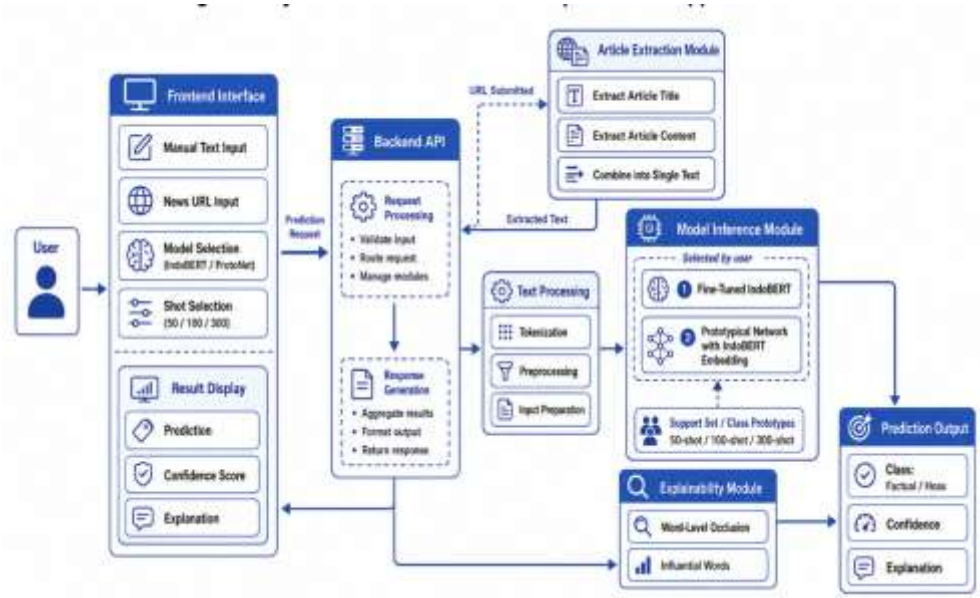

Figure 3. System architecture of the proposed web application

3.8 Explainable AI Method

Interpretability is achieved through a post-hoc word-level occlusion mechanism that quantifies feature sensitivity. The algorithm first records the baseline prediction probability output by the selected model for the unperturbed input text. Subsequently, individual lexical tokens are systematically omitted one at a time, and the modified text is re-evaluated by the inference pipeline. The absolute decrease between the baseline confidence score and the perturbed confidence score represents the impact score of each respective token. Tokens exhibiting the highest positive deviation are rendered in the visual interface as influential terms. While this sensitivity metric does not reconstruct internal Transformer attention dynamics, it provides intuitive local explanations regarding which lexical terms most heavily supported the classification outcome.

3.9 Experimental Setup

To establish an empirical benchmark, six distinct model configurations were systematically evaluated, as detailed in Table 5. The experimental design contrasts fine-tuned IndoBERT against ProtoNet across the 50-shot, 100-shot, and 300-shot sampling regimes. Performance evaluation is governed by standard classification metrics: accuracy, macro precision, macro recall, and macro F1-score, ensuring balanced performance assessment across both veracity classes.

Table 5. Experimental Model Configurations

Model	Shot Scenario	Main Mechanism	Evaluation Metrics
IndoBERT	50-shot	Supervised fine-tuning	Accuracy, Precision, Recall, F1-score
IndoBERT	100-shot	Supervised fine-tuning	Accuracy, Precision, Recall, F1-score
IndoBERT	300-shot	Supervised fine-tuning	Accuracy, Precision, Recall, F1-score
ProtoNet	50-shot	IndoBERT embedding + prototype distance	Accuracy, Precision, Recall, F1-score
ProtoNet	100-shot	IndoBERT embedding + prototype distance	Accuracy, Precision, Recall, F1-score

ProtoNet	300-shot	IndoBERT embedding + prototype distance	Accuracy, Precision, Recall, F1-score

The detailed hyperparameter settings governing model training and inference are summarized in Table 6. Both architectures share identical base encoder weights (indobenchmark/indobert-base-p1), input sequence limitations, epoch budgets, and learning rates. However, they diverge in optimization objectives, batch sizes, and model selection protocols, with IndoBERT utilizing validation-loss checkpoints and ProtoNet retaining final-epoch weights.

Table 6. Hyperparameter Configuration

Parameter	Fine-Tuned IndoBERT	ProtoNet
Pretrained encoder	indobenchmark/indobert-base-p1	indobenchmark/indobert-base-p1
Number of classes	2	2
Few-shot scenarios	50, 100, 300 per class	50, 100, 300 per class
Number of epochs	3	3
Learning rate	2×10^{-5}	2×10^{-5}
Training batch size	8	4
Evaluation/Test batch size	8	8
Loss function	Cross-entropy loss	Cross-entropy over negative distances
Representation	Classification output	CLS token embedding
Prototype calculation	Not applicable	Mean class embedding
Distance metric	Not applicable	Euclidean distance
Validation	Every epoch	Not separately applied
Model selection	Best validation macro F1-score	Final trained model
Random seed	42	42
Optimizer	AdamW (PyTorch fused)	Adam
Evaluation metrics	Accuracy, Macro Precision, Macro Recall, Macro F1-score	Accuracy, Macro Precision, Macro Recall, Macro F1-score

The computational framework and operational dependencies supporting the experimentation and system deployment are documented in Table 7. The core pipeline was developed in Python 3.12 utilizing PyTorch and Hugging Face Transformers, supported by FastAPI and Uvicorn for asynchronous backend serving.

Table 7. Software and Implementation Environment

Component	Specification
Programming language	Python 3.12
Deep-learning framework	PyTorch 2.10.0
Transformer library	Hugging Face Transformers 5.3.0
Tokenization library	Tokenizers 0.22.2
Numerical processing	NumPy 2.4.2
Backend framework	FastAPI 0.135.1
ASGI server	Uvicorn 0.41.0
ASGI framework	Starlette 0.52.1
Data validation	Pydantic 2.12.5
Model repository integration	Hugging Face Hub 1.5.0
Model serialization	Safetensors 0.7.0
Template engine	Jinja2 3.1.6
Training environment	Google Colab
Operating system	Linux 6.6.122+ (x86_64, glibc 2.35)
Processor	x86_64 CPU
GPU	Not used (CPU-only runtime)
Memory	12.67 GB RAM

By configuring the pipeline within a standard CPU-only environment with 12.67 GB of RAM, the experimental setup demonstrates that both fine-tuned IndoBERT and prototype-based inference can be executed and served effectively without reliance on specialized hardware accelerators. This standardized computational configuration guarantees deterministic execution across the evaluated shot configurations and establishes an

accessible benchmark for reproducing the detection pipelines and deploying the interactive web interface in practical, resource-constrained environments.

4. Results and Discussion

4.1 Results

4.1.1 Implementation Results

The primary engineering artifact developed in this research is an interactive web-based application designed to detect Indonesian fake news using deep contextual representations. The application architecture supports two distinct operational modalities: direct manual text submission and automated URL-based ingestion. Manual text entry allows users to paste unformatted news articles directly into the interface, whereas the URL extraction pipeline automatically scrapes and normalizes the headline and body text from remote web sources prior to classification. Users can dynamically select between the fine-tuned IndoBERT classifier and the IndoBERT-based Prototypical Network, alongside configuring the target few-shot training budget across 50-shot, 100-shot, or 300-shot scenarios. The primary user interface facilitating input submission and model parameter selection is shown in Figure 4.

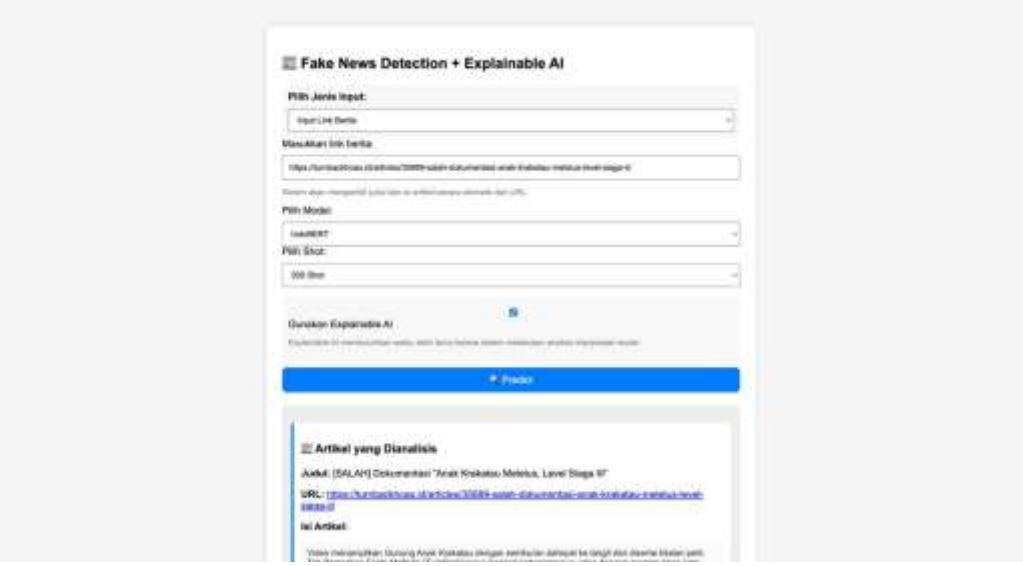

Figure 4. Main interface of the web application

Upon executing inference, the backend API returns a comprehensive diagnostic report comprising the predicted veracity class, the associated probability confidence score, the normalized input text, and visual token-level attributions. The resulting diagnostic view presented to the user is illustrated in Figure 5.

Figure 5. Prediction result page

These deployment outcomes verify that both parametric Transformer classifiers and non-parametric metric networks can be successfully operationalized within a responsive web architecture, transitioning experimental NLP models into accessible verification tools.

4.1.2 Fine-Tuned IndoBERT Results

The fine-tuned IndoBERT models were evaluated on the held-out testing partition containing 4,789 articles. Because this evaluation set remained completely isolated throughout training and hyperparameter checkpointing, it provides an unbiased quantification of model generalization on unseen Indonesian news texts. The quantitative performance metrics for IndoBERT across all evaluated shot configurations are documented in Table 8.

Table 8. Evaluation Results of Fine-Tuned IndoBERT

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score
IndoBERT 50-shot	0.9121	0.9118	0.9117	0.9117
IndoBERT 100-shot	0.9463	0.9469	0.9454	0.9460
IndoBERT 300-shot	0.9712	0.9715	0.9707	0.9710

The empirical results demonstrate robust classification capability across all sampling regimes. Even under the constrained 50-shot scenario, IndoBERT achieved an accuracy of 0.9121 and a macro F1-score of 0.9117. Expanding the supervision budget to 100-shot improved accuracy to 0.9463 and macro F1-score to 0.9460, culminating in the 300-shot setting with an accuracy of 0.9712 and a macro F1-score of 0.9710. This continuous performance trajectory highlights that the model efficiently leverages incremental labeled data to construct highly discriminative decision boundaries. Furthermore, the close alignment between precision and recall indicates balanced predictive capability without systematic bias toward either factual or deceptive articles. This strong baseline stems from the robust contextual representations established during IndoBERT pretraining on extensive Indonesian corpora (Koto *et al.*, 2020; Wilie *et al.*, 2020), which effectively transfer to nuanced news veracity classification.

4.1.3 Prototypical Network Results

The Prototypical Network was benchmarked against the identical held-out test partition. In this framework, the support instances define class prototype centroids within the IndoBERT embedding space, and query test instances are categorized based on their relative Euclidean proximity to these centroids. The empirical evaluation metrics for ProtoNet are summarized in Table 9.

Table 9. Evaluation Results of Prototypical Network

Model	Accuracy	Macro F1-score
ProtoNet 50-shot	0.9083	0.9082
ProtoNet 100-shot	0.9140	0.9138
ProtoNet 300-shot	0.9265	0.9263

ProtoNet demonstrated stable and consistent metric performance across all few-shot configurations, yielding an accuracy of 0.9083 and an F1-score of 0.9082 in the 50-shot scenario, progressing to 0.9140 accuracy and 0.9138 F1-score at 100-shot, and achieving 0.9265 accuracy and 0.9263 F1-score under the 300-shot regime. These findings confirm that distance-based prototype classification remains viable even when labeled instances are severely constrained. Increasing the support sample budget directly enhances the geometric stability and representativeness of the class centroids, reducing the sensitivity of prototype positions to individual sample outliers. The metric-learning mechanism of ProtoNet diverges fundamentally from supervised fine-tuning. Rather than optimizing a parameterized classification layer, ProtoNet relies on the intrinsic geometric clustering of the latent embedding space (Snell *et al.*, 2017). Because the classification decision is governed by metric distance, overall performance is strictly bound by the representational fidelity of the underlying encoder. Utilizing IndoBERT as the feature extractor ensures that rich syntactic and semantic dependencies are encoded prior to distance calculation, corroborating prior findings that encoder expressiveness is paramount in few-shot NLP tasks (Bao *et al.*, 2019; Dopierre *et al.*, 2021; Geng *et al.*, 2019).

4.1.4 Comparison Between IndoBERT and Prototypical Network

A comprehensive comparative assessment of fine-tuned IndoBERT and ProtoNet across all evaluation metrics is detailed in Table 10, with visual metric trajectories across sampling scenarios presented in Figure 6.

Table 10. Comparison of IndoBERT and Prototypical Network

Model	Shot Scenario	Accuracy	Macro Precision	Macro Recall	Macro F1-score
IndoBERT	50-shot	0.9121	0.9118	0.9117	0.9117
IndoBERT	100-shot	0.9463	0.9469	0.9454	0.9460
IndoBERT	300-shot	0.9712	0.9715	0.9707	0.9710
ProtoNet	50-shot	0.9083	0.9176	0.9156	0.9082
ProtoNet	100-shot	0.9140	0.9192	0.9185	0.9138
ProtoNet	300-shot	0.9265	0.9385	0.9386	0.9263

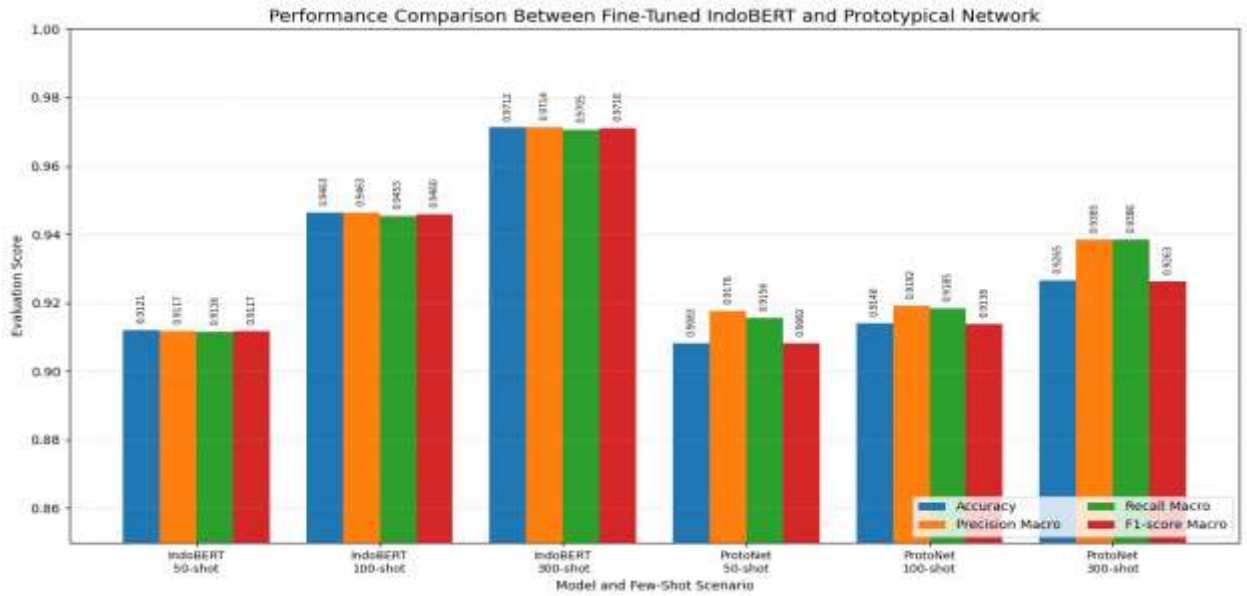

Figure 6. Accuracy, precision, recall, and F1-score comparison between fine-tuned IndoBERT and Prototypical Network across few-shot scenarios

While fine-tuned IndoBERT achieved superior overall performance across the broader evaluation, the competitive parity observed in the 50-shot configuration is particularly notable. In this highly constrained regime, IndoBERT attained an accuracy of 0.9121 and an F1-score of 0.9117, whereas ProtoNet achieved 0.9083 accuracy and 0.9082 F1-score. Notably, ProtoNet produced slightly higher macro precision (0.9176 vs. 0.9118) and macro recall (0.9156 vs. 0.9117). This indicates that when supervision is minimal, prototype averaging provides strong inductive regularization that mitigates parameter over-adaptation. As the labeled sample budget expands, the performance divergence between the two paradigms widens substantially. The macro F1-score margin favoring IndoBERT increased from 0.0035 in the 50-shot setting to 0.0322 in the 100-shot setting, reaching 0.0447 in the 300-shot configuration where IndoBERT established the peak benchmark of 0.9710 F1-score compared to 0.9263 for ProtoNet. This behavioral divergence reflects their underlying mechanics: supervised fine-tuning continuously updates Transformer attention matrices to fit task-specific decision boundaries, whereas ProtoNet relies on static metric distances whose prototype resolution exhibits diminishing marginal returns when intra-class variance remains wide.

4.1.5 External Article Testing

To evaluate model robustness in practical inference environments outside the curated Kaggle benchmark, real-world articles were scraped from BBC Indonesia (factual news) and TurnBackHoax.id (verified debunked misinformation). The classification predictions and associated confidence scores across all model configurations are documented in Table 11.

Table 11. External Article Testing Results

Source	Model	Shot	Prediction	Confidence
https://www.bbc.com/indonesia/articles/c75glvzgre3o	IndoBERT	50-shot	Factual	88.96%
https://www.bbc.com/indonesia/articles/c75glvzgre3o	IndoBERT	100-shot	Factual	95.64%
https://www.bbc.com/indonesia/articles/c75glvzgre3o	IndoBERT	300-shot	Factual	99.84%

https://turnbackhoax.id/articles/35689-salah-dokumentasi-anak-krakatau-meletus-level-siaga-iii	IndoBERT	50-shot	Hoax	72.13%
https://turnbackhoax.id/articles/35689-salah-dokumentasi-anak-krakatau-meletus-level-siaga-iii	IndoBERT	100-shot	Hoax	94.35%
https://turnbackhoax.id/articles/35689-salah-dokumentasi-anak-krakatau-meletus-level-siaga-iii	IndoBERT	300-shot	Hoax	99.85%
https://www.bbc.com/indonesia/articles/c75glvzgre3o	ProtoNet	50-shot	Factual	51.48%
https://www.bbc.com/indonesia/articles/c75glvzgre3o	ProtoNet	100-shot	Factual	51.35%
https://www.bbc.com/indonesia/articles/c75glvzgre3o	ProtoNet	300-shot	Factual	51.48%
https://turnbackhoax.id/articles/35689-salah-dokumentasi-anak-krakatau-meletus-level-siaga-iii	ProtoNet	50-shot	Hoax	51.55%
https://turnbackhoax.id/articles/35689-salah-dokumentasi-anak-krakatau-meletus-level-siaga-iii	ProtoNet	100-shot	Hoax	51.68%
https://turnbackhoax.id/articles/35689-salah-dokumentasi-anak-krakatau-meletus-level-siaga-iii	ProtoNet	300-shot	Hoax	51.77%

All twelve evaluated configurations correctly classified both external articles, verifying factual reporting for BBC Indonesia and deceptive framing for TurnBackHoax. However, probability confidence profiles differed sharply between architectures. IndoBERT exhibited sharp, highly confident predictions that scaled directly with shot budgets, reaching 99.84% and 99.85% confidence at 300-shot. Conversely, ProtoNet exhibited marginal probability differentials ranging narrowly from 51.35% to 51.77% across all shot settings. This marginal separation indicates that in metric space, out-of-distribution external texts reside near the equidistance threshold between class centroids. While these external tests provide encouraging preliminary evidence of system usability, they involve limited case samples and should be extended across broader news corpuses in future evaluations.

4.1.6 Explainable AI Results

The explainability engine visualizes prediction sensitivity by quantifying changes in class probability when individual tokens are systematically occluded. Terms whose deletion triggers the steepest drop in target class probability are visually highlighted, revealing the lexical anchors that primarily drove the inference. For instance, in investigative reporting, forensic, administrative, or legal terminology significantly stabilizes factual classification. A representative output of the word-level occlusion interface is shown in Figure 7.

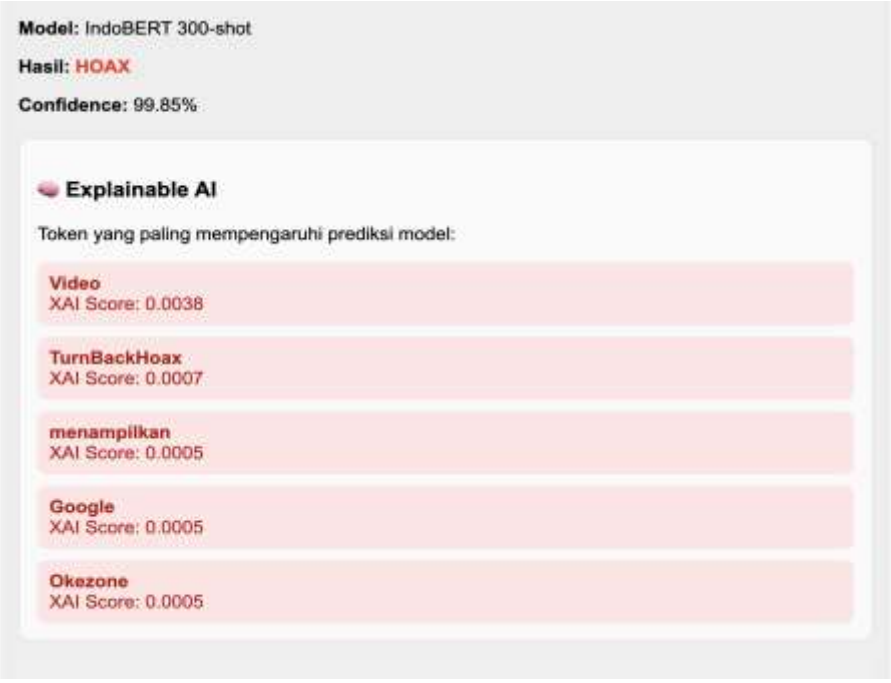

Figure 7. Explainable AI output using word-level occlusion

These attributions must be interpreted with methodological caution. Post-hoc token perturbation reflects input sensitivity rather than the complete internal reasoning dynamics of Transformer attention layers (Arrieta *et al.*, 2020; Rogers *et al.*, 2020). Consequently, highlighted terms serve as pragmatic decision-support signals for end users rather than absolute semantic proofs of veracity. Operational limitations remain, as single-token masking cannot capture multi-word colloquialisms or contextual negation, motivating future extensions into phrase-level attribution and entity-aware explainability.

4.2 Discussion

The empirical findings confirm that IndoBERT provides a powerful representational foundation for Indonesian fake news detection, aligning with prior evaluations of monolingual Transformer architectures (Koto *et al.*, 2020; Wilie *et al.*, 2020). Fine-tuned IndoBERT established the peak benchmark at 300-shot with 97.12% accuracy and a 97.10% macro F1-score, proving that targeted supervised adaptation efficiently learns discriminative journalistic patterns. Simultaneously, the ProtoNet implementation demonstrated the feasibility of metric-based few-shot learning, maintaining macro F1-scores above 0.908 across all settings without requiring extensive supervised head fine-tuning. From an engineering standpoint, the modular FastAPI deployment bridges algorithmic experimentation with real-world utility by supporting real-time web parsing and interpretable sensitivity heatmaps. Notwithstanding these positive results, several operational limitations warrant consideration. Model generalizability primarily stems from the specific linguistic distribution of the Kaggle source corpus, meaning performance may vary across specialized regional dialects or emerging news formats. This dependency is compounded by few-shot sampling sensitivity, wherein small support sets remain susceptible to instance-selection variance. Furthermore, automated URL extraction remains vulnerable to anti-scraping defenses, dynamic client-side DOM rendering, and paywall restrictions on commercial news portals. Additionally, token-level occlusion provides a localized proxy of model sensitivity rather than a holistic explanation of contextual narrative reasoning. Future research should implement multi-seed sampling evaluations, expand training corpora across cross-domain Indonesian news outlets, and advance explainability mechanisms toward structured phrase- and sentence-level semantic attribution.

5. Conclusion and Recommendations

This study successfully developed and evaluated an end-to-end Indonesian fake news detection web application integrating fine-tuned IndoBERT and an IndoBERT-based Prototypical Network (ProtoNet) across few-shot learning scenarios. The resulting platform supports dual input modalities through manual text submission and automated URL-based article extraction, while providing dynamic model selection, shot configuration adjustments, real-time confidence scoring, and word-level occlusion interpretability. By systematically evaluating parametric fine-tuning against non-parametric metric learning under 50-shot, 100-shot, and 300-shot regimes, this research establishes an empirical benchmark for Indonesian misinformation detection under varying levels of labeled data scarcity. The empirical findings demonstrate that fine-tuned IndoBERT achieved the strongest overall classification performance, attaining its peak result in the 300-shot setting with an accuracy of 0.9712 and a macro F1-score of 0.9710. The supervised fine-tuning mechanism proved highly effective at continuously adapting Transformer representations to task-specific journalistic patterns as the supervision budget expanded. Conversely, ProtoNet demonstrated remarkable resilience in the highly constrained 50-shot scenario, achieving a macro F1-score of 0.9082 along with slightly higher macro precision and recall than IndoBERT. These outcomes confirm that while supervised fine-tuning benefits most substantially from larger training samples, metric-based prototype estimation provides a viable and competitive classification strategy when labeled data are exceptionally scarce. The primary contribution of this research lies in unifying contextual Transformer fine-tuning, prototype-based few-shot learning, automated web content scraping, and post-hoc model interpretability within a single, cohesive software architecture. Furthermore, this work provides a rigorous comparative framework elucidating how monolingual Indonesian language embeddings behave when subjected to parametric classification versus geometric distance mapping in metric space. From a practical standpoint, the modular FastAPI deployment demonstrates that complex deep learning and explainability pipelines can be operationalized into accessible, interactive verification tools for end users without requiring local computational environments.

Despite these contributions, several operational limitations should be acknowledged. The classification models were evaluated on a single news corpus, meaning their cross-domain generalizability across specialized regional dialects, colloquial social media registers, or emerging misinformation formats warrants further investigation. In addition, the few-shot sampling protocol may introduce variance across different random splits, while the post-hoc word-level occlusion mechanism captures localized token sensitivity rather than the complete narrative and contextual reasoning of Transformer architectures. Operationally, the automated URL extraction pipeline remains susceptible to failure when encountering dynamic client-side rendering, anti-

scraping protections, or paywalled media platforms. To address these limitations, future research should expand evaluations across multi-seed few-shot sampling regimes and broader cross-domain Indonesian news corpora to verify distributional robustness. Subsequent studies could also benchmark alternative Transformer backbones, multilingual architectures, and hybrid graph-prototype networks to enhance metric separation in latent space. In addition, advancing explainable AI mechanisms toward phrase-level, sentence-level, and named-entity attribution will provide more contextually coherent explanations of deceptive framing. Finally, conducting formal empirical user-experience studies will be valuable to quantify end-user trust, interpretability comprehension, and decision-making accuracy when interacting with automated verification platforms.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Azizah, S. F. N., Cahyono, H. D., Sihwi, S. W., & Widiarto, W. (2023). Performance analysis of transformer based models (BERT, ALBERT and RoBERTa) in fake news detection. In *2023 6th International Conference on Information and Communications Technology (ICOIACT)* (pp. 425–430). IEEE. https://doi.org/10.1109/ICOIACT59844.2023.10455849

Bao, Y., Wu, M., Chang, S., & Barzilay, R. (2019). Few-shot text classification with distributional signatures. *arXiv preprint arXiv:1908.06039*. https://doi.org/10.48550/arXiv.1908.06039

Christian, Y., & Yap Rui Qi, K. O. (2022). Penerapan K-Means pada segmentasi pasar untuk riset pemasaran pada startup early stage dengan menggunakan CRISP-DM. *JURIKOM (Jurnal Riset Komputer)*, 9(4), 966–973. https://doi.org/10.30865/jurikom.v9i4.4486

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dopierre, T., Gravier, C., & Logerais, W. (2021). A neural few-shot text classification reality check. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 935–943). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.79

Erlina, E., Julyanto, J., Rustandi, J., Alexander, A., Francisco, L., Ma'muriyah, N., & Sabariman, S. (2023). Penerapan artificial intelligence pada aplikasi chatbot sebagai sistem pelayanan dan informasi online pada sekolah. *Journal of Information System and Technology*, 4(3), 610–621. https://doi.org/10.37253/joint.v4i3.6296

Fakhruzzaman, M. N., & Gunawan, S. W. (2021). Web-based application for detecting Indonesian clickbait headlines using IndoBERT. *arXiv preprint arXiv:2102.10601*. https://doi.org/10.48550/arXiv.2102.10601

Fathin, M. A., Sibaroni, Y., & Prasetyowati, S. S. (2024). Handling imbalance dataset on hoax Indonesian political news classification using IndoBERT and random sampling. *Jurnal Media Informatika Budidarma*, 8(1), 352–360. https://doi.org/10.30865/mib.v8i1.7099

Fawaid, J., Awalina, A., Krisnabayu, R. Y., & Yudistira, N. (2021). Indonesia's fake news detection using transformer network. In *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology* (pp. 247–251). Association for Computing Machinery. https://doi.org/10.1145/3479645.3479666

Geng, R., Li, B., Li, Y., Zhu, X., Jian, P., & Sun, J. (2019). Induction networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3904–3913). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1403

Han, L., Zhang, X., Zhou, Z., & Liu, Y. (2024). A multifaceted reasoning network for explainable fake news detection. *Information Processing & Management*, 61(6), Article 103822. https://doi.org/10.1016/j.ipm.2024.103822

Jocelynne, C., Wijayakusuma, I. G. N. L., & Harini, L. P. I. (2025). Detection of political hoax news using fine-tuning IndoBERT. *Journal of Applied Informatics and Computing*, 9(2), 354–360. https://doi.org/10.30871/jaic.v9i2.8989

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.66

Lei, S., Zhang, X., He, J., Chen, F., & Lu, C.-T. (2023). TART: Improved few-shot text classification using task-adaptive reference transformation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11014–11026). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.617

Liu, H., Wang, W., Li, H., & Li, H. (2024). TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 15556–15583). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.919

Liu, X., Gao, Y., Zong, L., & Xu, B. (2024). Improve meta-learning for few-shot text classification with all you can acquire from the tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 223–235). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.12

Lu, Y. J., & Li, C. T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 505–514). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.48

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.

Pekandi, L. A., Widjaja, R. G., Ananta, A., Harefa, J., & Jingga, K. (2025). Evaluating IndoBERT for Indonesian hoax news detection: A comparative study with ensemble and CNN-LSTM models. *Procedia Computer Science*, 269, 1625–1633. https://doi.org/10.1016/j.procs.2025.09.105

Prisscilya, V., & Girsang, A. S. (2024). Classification of Indonesia false news detection using BERTopic and IndoBERT. *Jurnal Indonesia Sosial Teknologi*, 5(8), 3913–3931. https://doi.org/10.59141/jist.v5i8.1310

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

Sehanobish, A., Kannan, K., Abraham, N., Das, A., & Odry, B. (2022). Meta-learning pathologies from radiology reports using variance aware prototypical networks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 332–347). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-industry.34

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

Simanjuntak, A., Lumbantoruan, R., Sianipar, K., Gultom, R., Simaremare, M., Situmeang, S., & Panggabean, E. (2024). Research and analysis of IndoBERT hyperparameter tuning in fake news detection. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 13(1), 60–67. https://doi.org/10.22146/jnteti.v13i1.8532

Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4077–4087). Curran Associates, Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3), 1–34. https://doi.org/10.1145/3386252

Wen, X., Tan, W., & Weber, R. (2025). GAProtoNet: A multi-head graph attention-based prototypical network for interpretable text classification. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 9891–9901). Association for Computational Linguistics.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 843–857). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.aacl-main.85

Yefferson, D. Y., Lawijaya, V., & Girsang, A. S. (2024). Hybrid model: IndoBERT and long short-term memory for detecting Indonesian hoax news. *IAES International Journal of Artificial Intelligence*, 13(2), 1913–1924. https://doi.org/10.11591/ijai.v13.i2.pp1913-1924

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 1–40. https://doi.org/10.1145/3395046