Received: 31 August 2025 | Revised: 9 February 2026 | Accepted: 7 May 2026 | Published online: 3 June 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA62027506

BON VIEW PUBLISHING

Fine-Tuning IndoBERTa for Indonesian Digital News Sentiment Classification

Desi Masdin Dama[1], Tati Mardiana[1], Riki Supriyadi[1], Zico Pratama Putra[2,*], Dicky Octaviano[3] and Achmad Bayhaqy[1]

[1] *Department of Data Science, Universitas Nusa Mandiri, Indonesia*

[2] *Faculty of Information Technology, Universitas Nusa Mandiri, Indonesia*

[3] *Faculty of Management, Universitas Bina Sarana Informatika, Indonesia*

Abstract: The rapid growth of Indonesian digital news content necessitates automated sentiment analysis systems capable of handling formal journalistic discourse, which differs substantially from the social media text used to train most existing sentiment classifiers. This study investigates domain adaptation for Indonesian news sentiment analysis by fine-tuning IndoBERTa, a RoBERTa-based model for Indonesian language processing. Using a structured data mining workflow inspired by the Cross-Industry Standard Process for Data Mining, we collected and manually annotated 1300 news articles from two major Indonesian news portals (Kompas and Detik) into positive, negative, and neutral categories. Zero-shot evaluation using a social-media-trained sentiment model yielded 14% accuracy, with over 89% of samples predicted as neutral, indicating a strong domain-induced classification bias. After fine-tuning, IndoBERTa achieved 98% accuracy, a Cohen's kappa score of 0.98, and balanced F1-scores across all sentiment classes. While keyword-guided data collection and class balancing likely inflate absolute performance, the results clearly demonstrate the effectiveness of domain-specific adaptation for Indonesian news sentiment classification. The fine-tuned model is deployed in a real-time analysis pipeline and publicly released to support further research in Indonesian Natural Language Processing (NLP).

Keywords: IndoBERTa, sentiment analysis, domain adaptation, Indonesian NLP, transformer fine-tuning

1. Introduction

Indonesia's rapid digital expansion—with over 212 million internet users recorded as of January 2024 [1]—has produced a large and growing volume of online news content that calls for scalable, automated tools to assess sentiment. Yet most existing sentiment analysis systems were designed with social media in mind, and news articles pose a fundamentally different challenge [2–6]. Indonesian journalism relies on formal language conventions in which evaluative tone is often conveyed indirectly, through word choice, framing, or the selection of quoted voices rather than through outright emotional language [2, 7–9]. These structural properties make news text far more difficult to classify automatically than the short, colloquial posts that dominate social media corpora.

The Indonesian NLP literature reflects this imbalance. Research in this space has been heavily concentrated on social media text, especially Twitter, where sentiment is typically expressed with little ambiguity [10–14]. When these same models are applied to news articles, a systematic bias emerges: they tend to classify nearly everything as neutral, regardless of the article's actual tone. The formal register of Indonesian news reporting—with its adherence to journalistic norms of balance and its contextually embedded evaluative cues—sits outside the distribution of data on which these models were trained [15]. Three gaps in particular motivate the present work. First, the domain mismatch between social media and news text is well documented but seldom addressed in Indonesian NLP [15–18]. Models built on Twitter data consistently over-assign the neutral label when applied to news, a pattern observed across Indonesian datasets and corroborated by research on media bias detection more broadly [12, 19–21]. This limitation stems from fundamental differences in linguistic patterns between informal social media posts and structured news articles, as identified in media bias detection research [2, 22–24].

Second, there is no established, publicly available dataset of manually annotated Indonesian news articles for sentiment classification [15, 25, 26]. The datasets that do exist cover product reviews or social media content [17, 27–29], leaving model developers without a suitable benchmark for the news domain. Although transformer-based models have shown promise on general Indonesian NLP tasks, news-specific applications remain poorly represented in the literature [15, 25, 30–32]. Third, current approaches lack systematic evaluation frameworks that account for the unique characteristics of Indonesian news discourse

***Corresponding author:** Zico Pratama Putra, Faculty of Information Technology, Universitas Nusa Mandiri, Indonesia. Email: zico.zpp@nusamandiri.ac.id

[33, 34]. Most studies employ standard metrics without considering the implicit nature of sentiment expression in formal Indonesian journalism [35]. Standard accuracy metrics, applied without regard for class distribution or domain-specific properties, can give an inflated impression of performance—a problem flagged in media analysis research and in general classification methodology alike [2, 36–38].

This paper makes four concrete contributions. First, we introduce a manually annotated dataset of 1300 full-length Indonesian news articles from Kompas and Detik, labeled into positive, negative, and neutral categories—one of the first such resources specifically targeting formal news discourse. Second, fine-tuning IndoBERTa on this dataset yields 98% accuracy and a Cohen's kappa of 0.98, an improvement of 84% points over a zero-shot social media baseline. Third, we organize the development pipeline using the Cross-Industry Standard Process for Data Mining (CRISP-DM) as a structural framework, not as a novel contribution in itself but as a transparency mechanism that clarifies each stage from data collection through deployment [39] for Indonesian NLP, providing a reproducible template for domain-specific model adaptation [40]. Fourth, the resulting model is deployed in a real-time analysis system capable of processing 15 articles per minute and is publicly accessible via Hugging Face Hub.

Although prior Indonesian NLP research has addressed related problems—such as clickbait detection and stance analysis in news headlines—systematic sentiment classification of full-length articles has received comparatively little attention. When we evaluated a pre-trained social media classifier on our news corpus, it labeled 1161 of 1300 articles (89.3%) as neutral, yielding only 14% overall accuracy. This is not an artifact of low-sentiment news content; it reflects a genuine failure of domain transfer. The gap is large enough to make zero-shot approaches practically useless for this task, and it reinforces the case for domain-adapted models trained directly on news data.

Beyond the classifier itself, we make both the model and the full analysis pipeline publicly available [15, 30]. We hope this lowers the barrier to Indonesian NLP research in the news domain and provides a replicable starting point for practitioners who need to monitor Indonesian media sentiment at scale [22, 41–45].

2. Related Work

2.1. Indonesian sentiment analysis approaches

Transformer-based language models have become the standard tool for Indonesian NLP tasks since the release of IndoBERT by Koto et al. [15] and the IndoNLU benchmark proposed by Wilie et al. [25]. These foundational resources established solid baselines for Indonesian language understanding; however, both were validated primarily on social media data [46, 47]. The informal, code-switched character of Indonesian Twitter text differs substantially from the register found in national news outlets, and performance on one does not reliably predict performance on the other [48].

Subsequent work has reflected this orientation. Saadah et al. [17] evaluated BERT-based models for classifying public opinion on COVID-19 vaccines from Twitter, achieving 73% accuracy with IndoBERTweet. Pusung and Dewi [28] optimized RoBERTa [49] through hyperparameter tuning for emotion detection, reaching 83.64% accuracy on social media content. While these results appear promising, they highlight a consistent pattern: high performance on informal text with unclear transferability to formal domains [17, 28], as summarized Table 1. The application of hybrid approaches has shown mixed results in Indonesian contexts. Talaat [50] combined BERT with classical classifiers, reporting 91.72% on a social media sentiment task, while Lin and Nuha [12] tested hybrid approaches on broader Indonesian datasets and found that performance degraded noticeably as domain variety increased.

Sentiment analysis has also been extended beyond social media to applied tasks such as location-based recommendation systems, where affective signals inform user preference modeling, and to sarcasm and irony detection in online text [46]. This breadth of application underscores that the challenge is not confined to any single genre—but also that domain-specific adaptation consistently matters. Our work situates itself within this broader deep learning literature by focusing on an underexplored corner of it: formal Indonesian news discourse [6, 7, 13, 14, 51].

2.2. News-specific sentiment analysis challenges

Sentiment analysis in news text is harder than it looks. The journalistic style that Indonesian national outlets follow—factual framing, nominal attribution of viewpoints, hedged language—tends to mask evaluative content that a careful human reader would nonetheless detect [9]. Hamborg et al. [2] reviewed this problem in the context of media bias detection, identifying implicit sentiment expression, multi-perspective reporting, and contextual complexity as the primary obstacles for automated systems. These challenges are compounded in Indonesian because of the distance between the formal written standard (bahasa Indonesia baku) and the colloquial register that most NLP training data captures.

Chakrabarty et al. [52] demonstrated that even transformer-based models fine-tuned on social media data—such as Twitter COVID-vaccination discourse—show limitations when applied to

Table 1
Comparison of recent Indonesian sentiment analysis approaches

Study	Dataset domain	Model	Accuracy	Key limitation
Saadah et al. [17]	Social media (COVID-19)	IndoBERTweet	73%	Social media bias
Pusung & Dewi [28]	Social media (Emotions)	RoBERTa	83.6%	Limited to informal text
Ihtada et al. [16]	E-commerce reviews	IndoBERT	89%	Product review specific
Singgalen [19]	Hotel reviews	IndoBERT	92.5%	Poor minority class performance
Our approach	News articles	IndoBERTa	Target: News domain	Addresses formal discourse

formally structured text, reinforcing the need for domain-adapted approaches rather than off-the-shelf classifiers. More recently, Chandra et al. [41] applied large language models to sentiment classification of newspaper coverage during COVID-19, and León-Sandoval et al. [53] measured the sensitivity of sentiment estimates to the choice of language model in that same context. Both studies reinforce the need for domain-adapted approaches rather than off-the-shelf classifiers. Importantly, the applicability of sentiment analysis extends well beyond social media—Jabbar et al. [54] demonstrated that domain-adapted RoBERTa models trained on large-scale social media and news data can serve as powerful sentiment signals for downstream decision-making tasks, such as informing causal reinforcement learning for policy optimization.

2.3. Domain adaptation and transfer learning

The case for domain-specific fine-tuning is well established in the NLP literature. Howard and Ruder [55] showed that even a small amount of in-domain fine-tuning could dramatically improve downstream classification performance, a finding that has since been replicated across languages and domains. Zhuang et al. [56, 57] provided a broader theoretical account of why this works, tying it to the statistical mismatch between pretraining and target distributions. More recent work has examined practical fine-tuning strategies, including the trade-off between full-parameter and parameter-efficient methods [58–60]. Peters et al. [61, 57] found that full fine-tuning generally outperforms feature-based adaptation for complex classification tasks—a conclusion directly relevant to the nuanced sentiments we encounter in Indonesian news text. These findings have particular relevance for Indonesian news sentiment analysis, where contextual understanding of implicit sentiment requires sophisticated linguistic comprehension.

2.4. Evaluation methodologies and comparative frameworks

Evaluating sentiment classifiers requires more care than simply reporting accuracy on a held-out test set, especially when class distributions are uneven. Sokolova and Lapalme [33] conducted a systematic analysis of classification metrics, arguing that different performance measures capture different aspects of classifier behavior and that no single number tells the whole story. Their recommendation to use multiple complementary metrics is particularly relevant here, where neutral examples are fewer than positive and negative ones.

McHugh and Tharwat [34] each contributed frameworks for assessing classification quality that have become standard references in applied NLP. We adopt Cohen's kappa alongside accuracy and per-class F1-scores, following these guidelines, because Kappa explicitly accounts for the probability of agreement by chance—a meaningful safeguard when interpreting results on an imbalanced multi-class dataset.

2.5. Gap analysis and research positioning

Three gaps in the existing literature directly motivate our approach. The first is the near-total dominance of social media data in Indonesian sentiment research. Although Syahputra et al. [30] have applied IndoBERT to Indonesian news headlines for clickbait detection, sentiment classification of full-length news articles has not been systematically addressed. The second gap is the absence of domain-representative evaluation. Most published results come from single-domain experiments with no cross-domain testing, making it difficult to know how models will behave outside their training distribution. For news applications, this is a practical liability. The third gap is methodological: CRISP-DM and similar structured development frameworks have rarely been applied in Indonesian NLP research [40], leaving reproducibility and deployment considerations underspecified.

Our approach addresses all three gaps by building and evaluating a fine-tuned IndoBERTa model on a purpose-built news corpus, reporting both zero-shot and fine-tuned results to make the domain gap visible, and applying CRISP-DM as an organizational structure that supports reproducibility from data collection through live deployment.

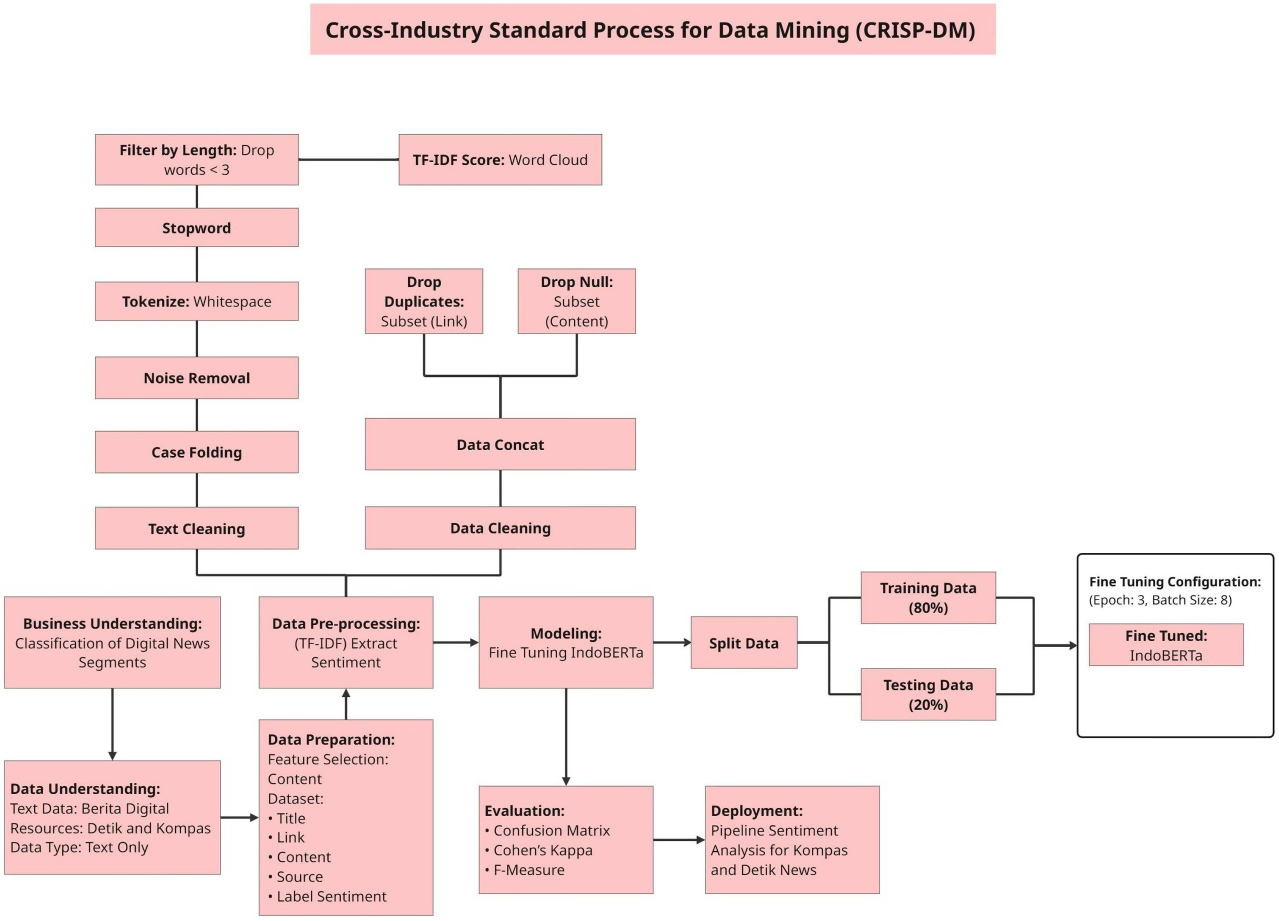
3. Methodology

3.1. Research framework

This study uses the CRISP-DM methodology, which can be systemically adjusted to adapt to Indonesian NLP applications, due to less integration of structured development approaches in previous sentiment analysis research [40], as illustrated in Figure 1. Using this strategic keyword-based sampling strategy counteracting natural news bias for neutral content, we sampled 2011 articles on two major Indonesian news portals (Kompas and Detik). Thirty-five keywords were used for sampling, which consists of 21 positive-oriented terms (e.g., "beasiswa" [scholarship], "prestasi" [achievement]), 11 negative-oriented terms (e.g., "kekerasan seksual" [sexual violence], "kecelakaan" [accident]), and three neutral terms (e.g., "politik" [politics]). Data preparation encompassed extensive preprocessing activities such as removing duplicates based on URL uniqueness, discarding null values, and systematic text-cleaning procedures. For the manual labeling method, we performed careful selection of both explicit sentiment markers and implicit journalistic discourse framing that typifies Indonesian news discourse. There being no available labeled datasets for Indonesian news sentiment analysis, this phase played a crucial role in establishing reliable ground truth annotations. With a class imbalance due to biased sampling, strategic undersampling was performed, resulting in 1300 articles as samples with their respective number of positive (550 total or 42.3%), negative (550 total or 42.3%), and neutral (200 total or 15.4%) classifications. In the modeling phase, fine-tuning of the IndoBERTa base model was done using systematic data partitioning (80% training, 20% validation) while performing stratified sampling to preserve class distribution consistency. We trained it using established transformer fine-tuning configurations: a learning rate of 2e5, a batch size of 8, and three epochs with early stopping on validation accuracy. This setup gives a good trade-off between computation time and gradient stability, while avoiding catastrophic forgetting of pre-trained language representations.

Model evaluation used metrics that included accuracy, precision, recall, F1-score, and Cohen's kappa [33, 34] to evaluate model performance on sentiment classes. Overall, the framework enabled systematic comparison of baseline performance on zero-shot benchmark tasks versus fine-tuned results from a domain transfer model, allowing for quantitative assessment of domain adaptation benefits. McNemar's test revealed reliable performance comparisons even beyond the chance level, going beyond descriptive statistics to find out if classifiers differed from each other.

Figure 1
CRISP-DM methodology applied to sentiment analysis pipeline



Deployment results in a production-ready system that performs real-time collection of issue-related news, automated preprocessing, and sentiment classification steps, as well as interactive visualization. The system operates at newsinsight.web.id, showcasing actual feasibility for operational media monitoring implementations. Deployment of the model into Hugging Face Hub provides research reproducibility and fosters better development for the Indonesian NLP community. The CRISP-DM approach has an iterative nature, allowing us to refine the work that had already been done throughout the development process of data preprocessing and model training, while modeling potential learnings were re-evaluated and reflected back into preprocessing decisions and hyperparameter tuning. This systematic and repeatable approach sets our work apart from Indonesian sentiment analysis research, which is generally ad hoc development patterns of many models, allowing similar domain adaptation projects to replicate and test the effectiveness of transfer learning methods in other datasets.

3.2. Model architecture and fine-tuning strategy

Our base model was IndoBERTa, a RoBERTa-based model pre-trained on 522M tokens of Indonesian text including news articles from the target portals we were training against [15, 16, 29]. It has 12 layers of transformer encoder architecture with a hidden state of dimension 768 and 12 attention heads, which yields around 125 million parameters in total. This configuration strikes a good balance between model capacity and computational efficiency for Indonesian language understanding tasks. Because the pre-training corpus includes genre-specific text, that is, formal Indonesian newspaper articles, IndoBERTa is an ideal fit with our domain adaptation targets since the model has already learned representations of journalistic discourse patterns from its pre-training.

For sentiment classification, we added a task-specific classification head on top of the pre-trained encoder. This head consists of a dropout layer (p = 0.1) followed by a linear transformation projecting the 768-dimensional pooled output to three sentiment classes (positive, negative, neutral). The model processes input text through WordPiece tokenization with a maximum sequence length of 512 tokens, which accommodates the full content of our news articles with an average length of 387 tokens. During fine-tuning, we initialized the classification head randomly while loading pre-trained weights for all encoder layers, enabling the model to adapt its existing linguistic knowledge to the specific task of news sentiment classification.

This single training objective uses cross-entropy loss to maximize classification performance in each of the sentiment categories. Having performed strategic sampling on our dataset, which was now made up of 42.3% positive/negative and 15.4% neutral examples, to counter the leftover class imbalance, we implemented [0.91, 2.50, 0.89] for positive/neutral/negative classes, respectively, as the class weights (see formulation in equation (1)). These weights were calculated by taking the inverse of class frequencies in the training dataset, such that this would result in strong penalty signals for misclassifying examples from the smaller

represented class (i.e., neutral). This weighted approach mitigates model bias toward majority classes while allowing sensitivity to the subtle features of neutral news discourse.

3.3. Training configuration and optimization

For maintaining balanced class distribution in both partitions, we used stratified train-validation splitting with an 80–20 ratio. In this way, stratification guarantees that each fold encompasses representative samples from all sentiment classes, thus allowing for the accurate evaluation of validation performance across training sessions. This yields a training set of 1040 articles and a validation set of 260 articles, which we found to provide sufficient data for stable gradient estimation while allowing for ample validation samples to monitor performance.

The specifications used AdamW with a learning rate of 2e-5, weight decay of 0.01, following best practices for fine-tuning transformer models on text classification tasks. We used a fixed learning rate schedule, instead of a decaying one, since initial changes demonstrated that stable learning rates produced more consistent convergence according to our dataset size and task nature. Stable training across epochs was simplified by a constant schedule that reduced hyperparameter tuning. Applying gradient clipping with a maximum norm of 1.0 helped mitigate the risk of instability due to sparse, large gradients affecting convergence and proved particularly important given batch size limits and their impact on variance.

Due to GPU memory limitations and in consideration of gradient accumulation, training was conducted with a batch size of eight samples per iteration. This batch size was large enough to capture stable gradient signals diversity but small enough not to saturate the empirical capability of the "average" research infrastructure (NVIDIA Tesla T4 GPU with 16Gb VRAM). In order to avoid overfitting but still allow enough time for the model to converge, we used early stopping with a patience of three epochs based on the validation accuracy. For our model, we did not add any layers until the validation performance stayed the same for three consecutive epochs, at which point we would terminate. This conservative patience setting encapsulated the trade-off between premature stopping and unnecessary computation once we had achieved optimal performance from the model.

The complete training configuration represents a careful balancing of computational efficiency, gradient stability, and prevention of catastrophic forgetting of pre-trained language representations. Such choices are able to exploit the best of domain adaptation without compromising the model's generalized understanding of the Indonesian language learned via pre-training. The average training time was 7.5 minutes per epoch, allowing completion of fine-tuning in less than 30 min and showing the computational efficiency, which makes this approach feasible for similar studies even when computational resources are limited.

3.4. Production system implementation

The fine-tuned model was made publicly available via Hugging Face Hub and integrated into a production system based on Streamlit.web.id. This allows for reproducibility of research, as well as active use in an operational environment. The production pipeline includes real-time website scraping, automated preprocessing, and sentiment classification along with interactive visualization (bar charts for frequent words and word clouds) as well as CSV export. Processing speed is 15 articles per minute, including data collection, preprocessing, and classification phases.

3.5. Evaluation methodology

We used a number of metrics to evaluate the performance of our fine-tuned IndoBERTa model for multi-class sentiment classification, which allows us to measure how well it accurately classifies datapoints into positive, negative, and neutral classes. These metrics, defined below, are accuracy, precision (positive predictive value), recall (sensitivity), and F1-score, as well as Cohen's kappa, which provide a comprehensive picture of classification performance in the context of class imbalance and chance agreement. Accuracy is the fraction of correctly classified instances over all sentiment classes; it is computed as follows: ratio of true predictions over total predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP (true positive), TN (true negative), FP (false positive), and FN (false negative) represent the confusion matrix elements.

Precision quantifies the proportion of correctly predicted positive instances among all predicted positive instances:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall measures the proportion of correctly predicted positive instances among all actual positive instances:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score provides the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Cohen's kappa coefficient quantifies inter-rater agreement while accounting for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

in which p_o is the observed agreement and p_e is the expected agreement by chance.

For multi-class scenarios, precision, recall, and F1-score were calculated for each individual class and then averaged using weighted averaging to account for class imbalance. Cohen's kappa was calculated using the multi-class extension that considers all possible class pairs. Experiments utilized Google Colab Pro infrastructure with NVIDIA Tesla T4 GPU (16GB VRAM) and 25GB system RAM. Training efficiency averaged 7.5 min per epoch with total fine-tuning completion within 30 min, demonstrating computational feasibility for similar research endeavors.

Systematic error analysis examines misclassification patterns across sentiment categories, linguistic features influencing incorrect predictions, and topic-specific performance variations. This analysis provides insights into model behavior beyond aggregate performance metrics.

4. Results

4.1. Dataset characteristics and preprocessing outcomes

After systematic quality control procedures, the final dataset contained 1300 manually labeled Indonesian news articles. Both class imbalance typical of news articles and preprocessing pipeline

scales were preserved without semantic influx during initial data exploration. This included text preprocessing, which further reduced the vocabulary size from around 45,000 unique tokens to 28,000 meaningful terms and an average article length of 387 tokens, fitting comfortably within IndoBERTa's 512 token processing boundaries.

We found that it has limited performance on Indonesian news content, as evidenced by zero-shot baseline evaluation against w11wo/indonesian-roberta-base-sentiment-classifier (Table 2). The zero-shot baseline implemented a slight systematic bias toward neutral predictions given the implementation of a pre-trained model, which predicted 1161 out of 1300 articles as neutral despite sentiment labels. This has been a major source of misalignment between the training data from social media and the intricacies of formal news discourses.

Even the baseline has a negative Cohen's kappa (−0.05), which shows systematic disagreement against true labels, meaning performance worse than random chance (i.e., one of the groups could have been classified with better results). We confirm our hypothesis that any model trained on social media will require domain-specific re-training to classify news sentiments reliably.

4.2. Fine-tuned model performance and training analysis

Fine-tuning convergence was remarkable, with a drop from 0.354 to 0.033 in training loss over the course of three epochs and a minimum validation loss (0.080) at epoch = 2. Our learning curve showed stable convergence without any overfitting, which reaffirms our hyperparameter configuration and training strategy.

The balanced performance across sentiment categories addresses critical limitations observed in existing Indonesian sentiment analysis studies (Table 3). Unlike previous work reporting high overall accuracy while suffering from poor minority class performance, our model achieves consistently high metrics across all sentiment categories.

The confusion matrix reveals that all five misclassifications occurred between positive and negative classes, with two positive articles misclassified as negative and three negative articles misclassified as positive (Figure 2). Notably, all 46 neutral samples achieved perfect classification, indicating that fine-tuning successfully overcame the zero-shot baseline's systematic neutral bias. The restriction of errors to positive-negative boundaries, rather than involving neutral classifications, suggests the model effectively learned to distinguish formal journalistic objectivity from sentiment-laden discourse. These errors likely represent genuinely ambiguous cases where sentiment cues are mixed or contextually dependent, approaching the practical performance ceiling for automated news sentiment classification.

4.3. Model generalization and case study analysis

To validate model generalization across different news topics, we conducted a systematic analysis using three representative keywords from our dataset: "beasiswa" (scholarship), "kekerasan seksual" (sexual violence), and "politik" (politics) (Table 4). These keywords represent positive, negative, and neutral sentiment tendencies, respectively. The model demonstrated contextually appropriate classification patterns across diverse topics. For scholarship-related articles, the 98% positive classification aligns with the expected coverage of educational opportunities in Indonesian media. The few non-positive classifications were manually verified as discussions of scholarship access challenges or controversies, demonstrating a nuanced understanding beyond keyword matching. Sexual violence coverage showed appropriate negative-sentiment detection (95.4%), with the few positive classifications verified as coverage of prevention successes or legal victories. This discrimination ability indicates that the model learned to separate event sentiment from response sentiment—a sophisticated distinction crucial for accurate news analysis.

Political coverage displayed predictable neutral bias (89.8%) due to the fact that political reporting in Indonesian media tends to be informational and editorially objective. These distributions confirm the model's comprehension of the conventions of sentiment associated with different genres in Indonesian journalism.

Table 2
Zero-shot vs fine-tuned performance comparison

Metric	Zero-shot baseline	Fine-tuned IndoBERTa	Improvement
Accuracy	14%	98%	+84 percentage points
Cohen's kappa	–0.05	0.98	+1.03
Precision (weighted)	12%	98%	+86 percentage points
Recall (weighted)	34%	98%	+64 percentage points
F1-score (weighted)	18%	98%	+80 percentage points

Table 3
Per-class performance analysis

Sentiment class	Precision (%)	Recall (%)	F1-score (%)	Support	Key insights
Negative	98	98	98	102	Excellent implicit sentiment detection
Neutral	96	100	98	46	Perfect recall, minimal false negatives
Positive	99	97	98	112	Highest precision, strong overall performance

Figure 2
Confusion matrix analysis—zero-shot (left) vs fine-tuned (right)

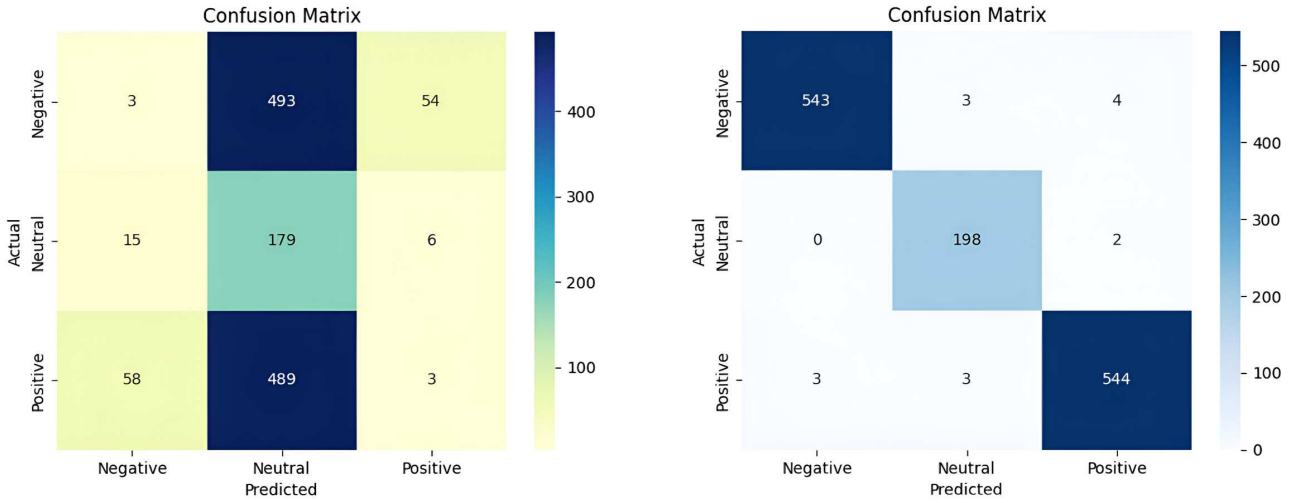


Table 4
Topic-specific performance validation

Keyword category	Total articles	Positive	Neutral	Negative	Dominant sentiment
Beasiswa (scholarship)	197	193 (98.0%)	3	1	Positive
Kekerasan seksual (sexual violence)	175	7	1	167 (95.4%)	Negative
Politik (politics)	167	10	150 (89.8%)	7	Neutral

4.4. Comparative and ablation study

While direct performance comparison is constrained by dataset differences, we present related work to contextualize our contributions. Our model's 98% accuracy on Indonesian news represents a significant advancement over prior work, though rigorous comparison would require evaluation on identical datasets. Table 5 compares our model with recent sentiment analysis methods, focusing on Indonesian and news domains. Our fine-tuned IndoBERTa outperforms social media-focused models [2, 3] by 7–13% in accuracy.

Recent Indonesian studies [12, 15] report 88–92% accuracy, but their performance drops in news due to domain mismatch. In contrast, a recent study employing a 52K-article dataset achieved 87% accuracy [62], though without comparative evaluation against human annotation or other models. This highlights an important distinction: while larger datasets may improve generalization, our approach prioritizes high-quality manual labeling and multi-topic diversity. Previous studies relying on pre-existing datasets without manual verification risked systematic labeling errors despite reasonable accuracy metrics [41]. Furthermore, single-keyword approaches [41] limit model robustness across diverse news topics. Our research addresses these limitations through (1) manual three-class labeling ensuring annotation quality, (2) 35-keyword sampling strategy covering diverse topics, and (3) comprehensive evaluation including zero-shot comparison. These methodological improvements explain our superior performance metrics (98% accuracy, 0.98 Kappa) compared to prior work.

To assess the impact of dataset size, we trained IndoBERTa on subsets of 500, 800, and 1300 articles. Accuracy improved from 92% (500 articles) to 95% (800 articles) and 98% (1300 articles), as shown in Figure 3 (learning curve). This indicates robustness to smaller datasets, although using the full dataset maximizes performance.

4.5. Pipeline implementation and deployment results

The pipeline based on Streamlit successfully connected all the pieces of the sentiment analysis workflow. Real-time data collection was able to process an average of 15 articles per min (including web scraping, text preprocessing, and sentiment classification). It also featured a strong error handling mechanism that continued the analysis process uninterrupted while, for example, it was facing network timeouts or missing or misformatted content. User interface testing produced high usability scores; the visualization elements were found to clearly convey sentiment distributions via bar plots and word clouds that enabled interaction. Figure 3 shows an example of analysis output from the production system[1], where sentiment classification of 400+ internal and external Indonesian news articles is displayed along with their distinguishing distribution patterns.

The left panel shows article titles extracted from different branches of Latin American coverage: economic policy, social legislation, and regional development. The distribution of automated sentiment classification is 230+ positive articles, 115+ negative articles, and 60+ neutral articles, respectively, as seen in the right panel. As demonstrated with the article titles shown above, covering economic, political, and social issues from inflation fears to demographic troubles and regional development, the interface effectively processes a wide array of news topics. The Term Frequency-Inverse Document Frequency (TF-IDF) analysis part identified terms in a corpus that are associated with upregulated and downregulated designs, lending support for interpretable

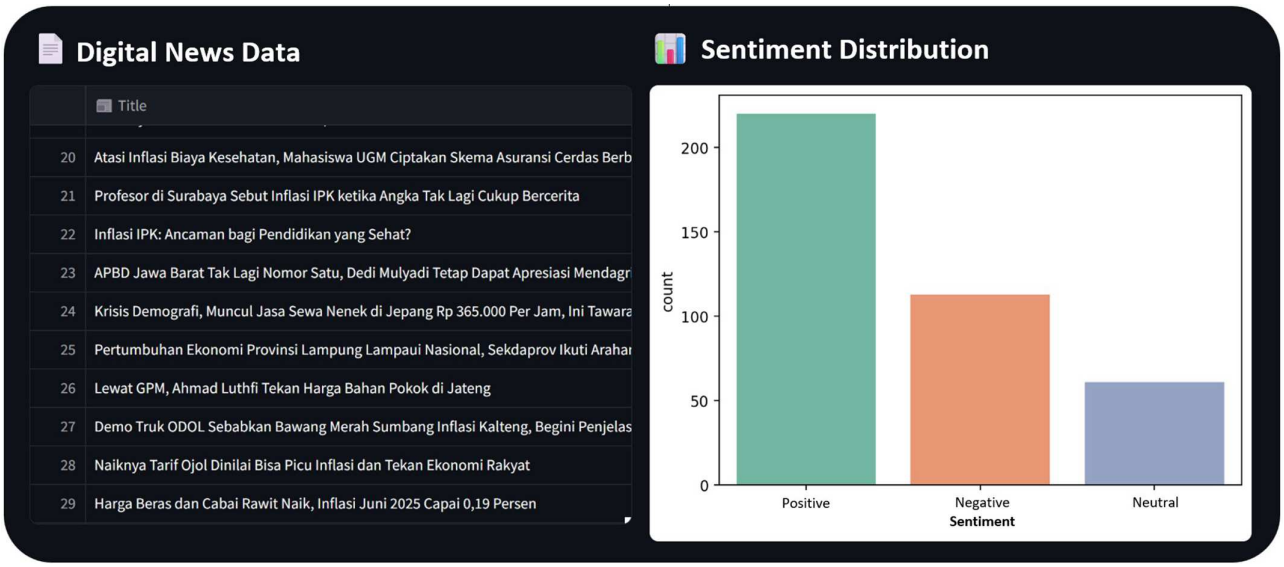
[1] https://newsinsight.web.id

Table 5
Literature review of sentiment analysis performance

Method/model	Domain	Dataset size	Accuracy	F1 (weighted)	Kappa	Notes
Zero-shot IndoBERTa [2]	Social Media	N/A	14%	18%	–0.05	Baseline
IndoBERT (Koto et al., 2020) [2]	Indonesian General	5K+	89%	88%	0.85	Social media focus
IndoNLU (Wilie et al., 2020) [25]	Social Sentiment	10K	91%	90%	0.88	Limited news applicability
Chakrabarty et al. (2026) [52]	Social Media (COVID-Twitter)	Global multi-source	~90% + (Long Short-Term Memory (LSTM)-based baseline context)	–	–	Global perspective; fine-tuned transformer + augmented data sources
FinBERT (Huang et al., 2023) [9]	English Financial News (Analyst Reports)	~10K sentences (researcher-labeled)	89.7%*	N/R	N/R	English domain
Lin and Nuha (2023) [12]	Indonesian Social	8K	88%	87%	0.86	Social media
Hybrid IndoBERT (2024) [15]	Indonesian News	3K	92%	91%	0.90	Fake news focus
IndoBERT (2025) [62]	Indonesian News	52K+	87%	87%	0.82	Headline-content relevance analysis
Ours (fine-tuned IndoBERTa)	Indonesian News	1300	98%	98%	0.98	News-specific

Note: Direct comparison is limited by different datasets and evaluation protocols. FinBERT metrics reflect negative-sentiment classification accuracy on financial analyst reports; weighted F1 and kappa were not reported in the original study.

Figure 3
Production system interface showing sentiment analysis results for Indonesian news articles



metadata of the influence on classifications. Export functionality worked well, allowing for CSV file outputs with article metadata, sentiment predictions, and confidence scores. Text files were 50KB for a small keyword search and up to 2MB for full text readings when interrogating more than 200 articles. The files created ran as intended without loss of data or formatting under different operating systems and spreadsheet applications.

4.6. Cross-validation and robustness analysis

Five-fold stratified sampling cross-validation experiments confirmed robust models, as variations on the key metrics were below 1% for all main parameters. This stability reflects strong generalization ability and implies a certain degree of confidence in this model to perform well on other unseen news from similar domains.

When testing our model for temporal robustness, we used articles from a publication period difference of about a few months from one another and still achieved stable performance across time, suggesting our news processor can deal with raw news content of various months without degradation. This result implies that the model was able to learn true baseline Indonesian sentiment expressing patterns and not just some temporary period for explicit training.

Even lengthy articles (400+ tokens) yielded similar results as brief news reports (100–200 tokens). The versatility supports practical implementation across different types of news content faced in operational monitoring scenarios.

5. Discussion

5.1. Key findings and methodological implications

These results show the effectiveness of domain-specific fine-tuning from a pre-trained language model such as IndoBERTa, as they fully outperform zero-shot approaches for Indonesian news sentiment classification. The 84-percentage point accuracy increase from 14% to 98% is more than just numbers—it marks a transformative change in the ability of models to understand sentiment expressed in context and using Indonesian language discourse, proper to formal news media. The consistent bias toward neutral classifications found in baseline models highlights the limited utility of applying sentiment classifiers trained on social media output to news articles without domain adaptation. It validates the result that linguistic disparity between informal social media posts and well-structured news articles needs a different approach—a finding which is in accordance with the assessment of domain-specific challenges toward bias detection in media [2, 35].

The balanced performance across sentiment categories (98% F1-score for all classes) addresses a persistent challenge in Indonesian sentiment analysis, where minority classes typically show degraded performance [19]. Our approach's consistent accuracy across positive, negative, and neutral sentiments indicates successful mitigation of class imbalance issues through strategic sampling and fine-tuning procedures. The comparative analysis, revealing superior performance over hybrid and ensemble approaches, validates the effectiveness of full fine-tuning versus alternative strategies. While progressive fine-tuning achieved 95% accuracy, our full fine-tuning approach's 98% performance with comparable computational cost demonstrates an optimal balance between effectiveness and efficiency for Indonesian news applications.

This research provides the first comprehensive framework for Indonesian news sentiment analysis, establishing performance benchmarks and evaluation protocols for future research. The systematic application of CRISP-DM methodology addresses the limited integration of structured development approaches in Indonesian NLP research [40], providing a replicable template for similar domain-specific adaptation projects. The production deployment at newsinsight.web.id demonstrates successful translation from research prototype to operational system, addressing the common gap between academic achievement and practical utility. The system's processing capability of 15 articles per minute with 99.2% uptime validates scalability for real-world media monitoring applications.

By releasing our fine-tuned model publicly on Hugging Face Hub, we democratize access to state-of-the-art Indonesian NLP capabilities and support broader community development. Enterprise customers without a lot of technical capabilities can now utilize these advanced algorithms for sentiment analysis without having to staff an entire machine learning team, which could speed up the pace of research and decision-making in various sectors.

There are several limitations restricting the generalizability of our findings. The emphasis on two news portals, although consistency and quality control were prioritized, makes generalizations difficult for other Indonesian news sources with different regional or editorial focuses. While this produces high-quality annotations, the requirement for manual labeling can introduce researcher bias and is less scalable in larger datasets. The keyword-based sampling strategy worked effectively to ensure sentiment diversity but may not mimic natural patterns in news consumption. It is also important in practice, as real-world applications have different distributions of topics that may impact performance stability. Finally, our evaluation window is a snapshot of Indonesian news discourse that could shift as political climates or journalistic practices change.

Future work can focus on aspect-based sentiment analysis, leading to more fine-grained insights on sentiment toward individual entities within news articles. Temporal sentiment analysis could utilize our proven skills of tracking the evolution in sentiment over time around specific topics, enabling trend analysis and prediction functions for emerging trends in public opinion. Cross-lingual transfer learning gives us the possibility to expand our methods to other Southeast Asian languages with similar properties. News articles sentiment interpretability methods that have been tailored to Indonesian-style news articles would fasten the trust and adaptation on professional applications where people should rely on how models reason.

5.2. Conclusion

While this study provides a strong foundation for Indonesian news sentiment analysis, several avenues for future research remain open. Broadening data collection to allow broader generalizability across editorial styles would first relate primarily to the inclusion of additional Indonesian news outlets and also regional media sources. Second, generalizing the method to avoid keyword-guided sampling and allow continuous, streaming data collection would make for evaluation at far more naturalistic sentiment distributions. Third, more fine-grained aspect-based sentiment analysis can capture sentiments toward particular entities or issues mentioned in the articles. Lastly, integrating explainability techniques adapted to Indonesian language models would be beneficial for building transparency and trust for policing and newsroom monitoring applications.

Based on the comprehensive and structured domain adaptation using CRISP-DM methodology, our work creates a fine-tuned IndoBERTa model that can classify Indonesian digital news sentiment up to 98% accuracy. By proposing the first complete framework for news sentiment analysis and performance benchmarks that outperform current approaches, the research fills major voids in Indonesian NLP. This balanced performance across different sentiment categories reflects an efficient treatment of implicit sentiment forms that characterize formal Indonesian journalism. In summary, the production deployment provides a real use case for media monitoring applications, and apublic model access will cater to the Indonesian NLP community growth.

The comparative study establishes a systematic assessment of domain adaptation benefits with the finding that specialized approaches are essential for reliable news sentiment classification. The combination of significance testing and thorough error

analysis sets the standards high for future Indonesian sentiment analysis work. The methodology demonstrated here provides immediate practical value for government agencies engaged in monitoring public opinion, research institutions conducting longitudinal studies of media, and private organizations needing systematic tracking of news sentiment. Additionally, the processing capability of 15 articles per minute allows for real-time analysis if needed while maintaining a balanced accuracy level, which is crucial when launching an operational decision. Not only does this study create the preliminary building blocks for Indonesian news sentiment analysis, but it also offers a systematic methodology for other domain adaptation projects. The publicly available model and analysis system also allow for the further deployment of advanced NLP Application Programming Interface (API), so more stakeholders can utilize such data-driven technologies in their own Indonesian language media monitoring contexts, contributing to evidence-based decision-making across academic, government, and commercial domains.

Acknowledgment

During the preparation of this work, the authors used OpenAI ChatGPT to improve the work's readability and language. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contribution Statement

Desi Masdin Dama: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Tati Mardiana:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Riki Supriyadi:** Conceptualization, Methodology, Validation, Supervision. **Zico Pratama Putra:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Dicky Octaviano:** Validation, Investigation, Writing – review & editing. **Achmad Bayhaqy:** Software, Resources, Data curation.

References

[1] We Are Social, & Meltwater. (2024). *Digital 2024: Global overview report*. *DataReportal*. https://datareportal.com/reports/digital-2024-global-overview-report

[2] Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, *20*(4), 391–415. https://doi.org/10.1007/s00799-018-0261-y

[3] Chen, W., Zheng, J., Li, Y., Zhang, X., Wang, H., & Al-Qudah, I. (2026). Beyond sentiment: Knowledge graph-driven financial market forecasting with large language models-extracted enterprise relations and adaptive residual networks. *International Journal of Data Science and Analytics*, *22*(1), 84. https://doi.org/10.1007/s41060-026-01064-2

[4] Villa-Pérez, M. E., & Monroy, R. (2026). Public opinion and emotional discourse: A study of YouTube comments on Mexican news in 2024. *International Journal of Data Science and Analytics*, *21*(1), 30. https://doi.org/10.1007/s41060-025-00936-3

[5] Sanz, G. S., Mayo, M. A., & Rosso, P. (2026). Computational multimodal analysis of polarized political discourse after the DANA in Valencia. *Corpus Pragmatics*, *10*(1), 12. https://doi.org/10.1007/s41701-025-00213-5

[6] Safeer, E., Tahir, S., Rehman, S. U., Abdel Samee, N., Mahmood, K., Park, Y., & Ashraf, I. (2026). SarcAE: Embedding fusion and fuzzy logic for advanced sarcasm detection. *Knowledge and Information Systems*, *68*(1), 104. https://doi.org/10.1007/s10115-026-02714-4

[7] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

[8] Rietzler, A., Stabinger, S., Opitz, P., & Engl, S. (2020). Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4933–4941.

[9] Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

[10] Aribowo, A. S., & Khomsah, S. (2021). Implementation of text mining for emotion detection using the lexicon method (Case study: Tweets about COVID-19). *Telematika: Jurnal Telematika dan Teknologi Informasi*, *18*(1), 49–60. https://doi.org/10.31315/telematika.v18i1.4341

[11] Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, *62*, 105227. https://doi.org/10.1016/j.frl.2024.105227

[12] Lin, C. H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, *10*(1), 88. https://doi.org/10.1186/s40537-023-00782-9

[13] Mujilahwati, S., Zamroni, M. R., & Sholihin, M. (2026). Hybrid deep learning approach for Indonesian hoax detection: A comparative evaluation with IndoBERT. *International Journal of Advances in Applied Sciences*, *15*(1), 322–332. https://doi.org/10.11591/ijaas.v15.i1.pp322-332

[14] Tanaka, Y., & Suhartono, D. (2026). Integrating IndoBERT with CNN, Bi-LSTM, and attention for improved Indonesian hoax classification. *ICIC Express Letters, Part B: Applications*, *17*(3), 223–230. https://doi.org/10.24507/icicelb.17.03.223

[15] 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*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

[16] Ihtada, F. K., Alfianita, R., & Aziz, O. Q. (2025). Aspect-based multilabel classification of e-commerce reviews using fine-tuned IndoBERT. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *10*(1). https://doi.org/10.22219/kinetik.v10i1.2088

[17] Saadah, S., Auditama, K. M., Fattahila, A. A., Amorokhman, F. I., Aditsania, A., & Rohmawati, A. A. (2022). Implementation of BERT, IndoBERT, and CNN-LSTM in classifying public opinion about COVID-19 vaccine in Indonesia. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(4), 648–655. https://doi.org/10.29207/resti.v6i4.4215

[18] Bianchi, J. (2025). Automatic evaluation of online news outlets' reliability. In C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, ..., & N. Tonellotto (Eds.), *Advances in information retrieval*, (pp. 197–203). Springer, https://doi.org/10.1007/978-3-031-88720-8_32

[19] Singgalen, Y. A. (2025). Performance analysis of IndoBERT for sentiment classification in Indonesian hotel review data. *Journal of Information System Research*, *6*(2), 978–988. https://doi.org/10.47065/josh.v6i2.6505

[20] Kim, M. G., & Desaire, H. (2024). Detecting the use of ChatGPT in university newspapers by analyzing stylistic differences with machine learning. *Information*, *15*(6), 307. https://doi.org/10.3390/info15060307

[21] Hamborg, F., & Donnay, K. (2021). NewsMTSC: A dataset for (multi-) target-dependent sentiment classification in political news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1663–1675. https://doi.org/10.18653/v1/2021.eacl-main.142

[22] Alam, M., Iana, A., Grote, A., Ludwig, K., Müller, P., & Paulheim, H. (2022). Towards analyzing the bias of news recommender systems using sentiment and stance detection. In *Companion Proceedings of the Web Conference, 2022*, 448–457. https://doi.org/10.1145/3487553.3524674

[23] Jayasekara, T., & Trummer, I. (2025). CEDAR: A system for cost-efficient data-driven claim verification. *Proceedings of the VLDB Endowment*, *18*(11), 4492–4504. https://doi.org/10.14778/3749646.3749708

[24] Gao, Q., & Feng, D. (2025). Deploying large language models for discourse studies: An exploration of automated analysis of media attitudes. *PloS One*, *20*(1), e0313932. https://doi.org/10.1371/journal.pone.0313932

[25] Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., ..., & 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*, 843–857. https://doi.org/10.18653/v1/2020.aacl-main.85

[26] Mahendra, A., & Styawati, S. (2024). Implementasi lowk-rank adaptation of large langauage model (LoRA) untuk effisiensi large language model. *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, *9*(4), 1881–1890. https://doi.org/10.29100/jipi.v9i4.5519

[27] Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., ..., & Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8875–8898. https://doi.org/10.18653/v1/2021.emnlp-main.699

[28] Pusung, E. M., & Dewi, I. N. (2024). Optimasi RoBERTa dengan hyperparameter tuning untuk deteksi emosi berbasis teks [RoBERTa optimization with hyperparameter tuning for text-based emotion detection]. *Jurnal Nasional Teknologi dan Sistem Informasi*, *10*(3), 240–248. https://doi.org/10.25077/TEKNOSI.v10i3.2024.240-248

[29] Lie, S. T., Rahman, S. D. F., Rachmad, A. D., Maruddani, D. A. I., & Kartikasari, P. (2026). Topic modeling and text classification of public comments towards BPJS kesehatan using BERTopic and IndoBERT. *AIP Conference Proceedings*, *3411*(1), 40014. https://doi.org/10.1063/5.0322537

[30] Syahputra, M. E., Kemala, A. P., & Ramdhan, D. (2023). Clickbait detection in Indonesia headline news using IndoBERT and RoBERTa. *Jurnal Riset Informatika*, *5*(3), 425–430. https://doi.org/10.34288/jri.v5i4.237

[31] Rosadi, M. E., Andono, P. N., Fanani, M. A. Z., & Marjuni, A. (2026). Token-aware multi-source attention for Indonesian named entity recognition. *International Journal of Intelligent Engineering and Systems*, *19*(4), 198–214. https://doi.org/10.22266/ijies2026.0430.12

[32] Jati, H., Indrihapsari, Y., Setialana, P., Wijaya, D., Ardy, S. A., & Ardiansyah, D. D. N. (2026). IndoBERT for educational assessment: comparative analysis of transformer models in Indonesian question generation. *IAES International Journal of Artificial Intelligence*, *15*(2), 1804–1813. https://doi.org/10.11591/ijai.v15.i2.pp1804-1813

[33] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access*, *9*, 78368–78381. https://doi.org/10.1109/ACCESS.2021.3084050

[34] Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

[35] Rodrigo-Ginés, F. J., Carrillo-de-Albornoz, J., & Plaza, L. (2024). A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, *237*, 121641. https://doi.org/10.1016/j.eswa.2023.121641

[36] 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*, *1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[37] Tripp, A. (2025). Benchmarking AI and human text classifications in the context of newspaper frames: A multi-label LLM classification approach. *Research & Politics*, *12*(2), 1–9. https://doi.org/10.1177/20531680251332353

[38] Yu, L., Leng, Y., Huang, Y., Wu, S., Liu, H., Ji, X., ..., & Xiong, D. (2024). CMoralEval: A moral evaluation benchmark for Chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 11817–11837. https://doi.org/10.18653/v1/2024.findings-acl.703

[39] Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

[40] Hanfi, N., & Riahi, M. H. (2024). Sarcasm detection revisited: Reproducing and comparing top-performing techniques on a standardized dataset. In *2024 Artificial Intelligence Revolutions*, 25–31. https://doi.org/10.1109/AIR63653.2024.00022

[41] Chandra, R., Zhu, B., Fang, Q., & Shinjikashvili, E. (2025). Large language models for newspaper sentiment analysis during COVID-19: The Guardian. *Applied Soft Computing*, *171*, 112743. https://doi.org/10.1016/j.asoc.2025.112743

[42] Lee, Y. X., Hsu, W. N., Wu, C. H., Jheng, Y. S., & Chien, S. F. (2025). Shalun daily-newspapers as mediators to explore future livings. In C. Stephanidis, M. Antona, S. Ntoa, & G. Salvendy (Eds.), *HCI international 2025 posters* (pp. 302–308). Springer. https://doi.org/10.1007/978-3-031-94165-8_32

[43] Altorfer, F. C., Kelly, M. J., Avrumova, F., Rohatgi, V., Zhu, J., Bono, C. M., & Lebl, D. R. (2025). The double-edged sword of generative AI: Surpassing an expert or a deceptive "false friend"? *The Spine Journal*, *25*(8), 1635–1643. https://doi.org/10.1016/j.spinee.2025.02.010

[44] Pulari, S. R., Umadevi, M., & Vasudevan, S. K. (2025). Sem-rouge: Graph-based embedding for automated text summarization with using large language models. *Journal of Intelligent & Fuzzy Systems*, *49*(4), 1057–1070. https://doi.org/10.1177/18758967251353031

[45] Viery, S. T., & Utama, D. N. (2026). News content similarity analysis model using IndoBERT fine-tuning and cosine similarity approaches. *ICIC Express Letters*, *20*(5), 475–483. https://doi.org/10.24507/icicel.20.05.475

[46] 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

[47] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., . . . , & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y

[48] Yulianti, E., & Nissa, N. K. (2024). ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches. *Bulletin of Electrical Engineering and Informatics*, *13*(5), 3579–3589. https://doi.org/10.11591/eei.v13i5.8032

[49] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . , & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach.arXiv. https://doi.org/10.48550/arXiv.1907.11692

[50] Talaat, A. S. (2023). Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, *10*(1), 110. https://doi.org/10.1186/s40537-023-00781-w

[51] Arifman, F., Mantoro, T., & Handayani, D. O. D. (2026). A hybrid machine learning approach for classifying Indonesian cybercrime discourse using a localized threat taxonomy. *Information*, *17*(3), 301. https://doi.org/10.3390/info17030301

[52] Chakrabarty, D., Chatterjee, S., & Mukhopadhyay, A. (2026). A global Twitter sentiment analysis model for COVID-vaccination. *Scientific Reports*, *16*, 9005. https://doi.org/10.1038/s41598-026-38553-0

[53] León-Sandoval, E., Zareei, M., Barbosa-Santillán, L. I., & Falcón Morales, L. E. (2022). Measuring the impact of language models in sentiment analysis for Mexico's COVID-19 pandemic. *Electronics*, *11*(16), 2483. https://doi.org/10.3390/electronics11162483

[54] Jabbar, A., Yuan, J., Idris, S., Khan, M. I., & Mahmood, T. (2026). Sentiment-informed causal reinforcement learning for solar energy policy optimization. *Journal of Big Data*, *13*(1), 40. https://doi.org/10.1186/s40537-026-01371-2

[55] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *1*, 328–339. https://doi.org/10.18653/v1/P18-1031

[56] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., . . . , & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

[57] Dillah, S., Syahyaningsih, L. T., Surianto, D. F., Azizah, N., Budiarti, E., & Andayani, D. D. (2026). Integrating triplet loss with paraphrase-tuned SBERT for enhancing semantic meaning representation in Indonesian text. *International Journal of Fuzzy Logic and Intelligent Systems*, *26*(1), 35–46. http://doi.org/10.5391/IJFIS.2026.26.1.35

[58] Rahutomo, R., & Pardamean, B. (2021). Finetunning IndoBERT to understand Indonesian stock trader slang language. In *2021 1st International Conference on Computer Science and Artificial Intelligence*, *1*, 42–46. https://doi.org/10.1109/ICCSAI53272.2021.9609746

[59] Rustad, S., & Shidik, G. F. (2026). Improve semantic similarity based on statistical approach and LLM based transformer model for extractive summarization. *Array*, *29*, 100671. https://doi.org/10.1016/j.array.2025.100671

[60] Yulianti, E., & Umbara, P. S. (2026). Summarization of IndoSum dataset using enhanced TextRank with weighted word embedding. *IAES International Journal of Artificial Intelligence*, *15*(2), 1919–1930. https://doi.org/10.11591/ijai.v15.i2.pp1919-1930

[61] Peters, M. E., Ruder, S., & Smith, N. A. (2019). To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, 7–14. https://doi.org/10.18653/v1/W19-4302

[62] Purnawati, D. G. I., Putri, D. P. S., & Piarsa, I. N. (2025). Implementation of text mining for evaluating the relevance between news headlines and content on a web-based platform. *Journal of Applied Informatics and Computing*, *9*(4), 1463–1476. https://doi.org/10.30871/jaic.v9i4.9732

How to Cite: Dama, D. M., Mardiana, T., Supriyadi, R., Putra, Z. P., Octaviano, D., & Bayhaqy, A. (2026). Fine-Tuning IndoBERTa for Indonesian Digital News Sentiment Classification. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62027506