Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229−2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

Explainable Aspect-Based Sentiment Analysis with Contrast-Aware IndoBERT for Indonesian Public Service Reviews

Muhammad Shihab Fathurrahman Jondien*, Taqwa Hariguna, Dhanar Intan Surya Saputra

Magister of Computer Science, Amikom Purwokerto University, Purwokerto, Indonesia
Email: [1,*]fathurshihab@gmail.com, [2]taqwa@amikompurwokerto.ac.id, [3]dhanarsaputra@amikompurwokerto.ac.id
Correspondence Author Email: fathurshihab@gmail.com
Submitted: **09/01/2026**; Accepted: **05/03/2026**; Published**: 06/03/2026**

Abstract−This study presents an Explainable IndoBERT with Contrast-Aware Attention framework for Aspect-Based Sentiment Analysis (ABSA) on Indonesian public service reviews. The proposed model integrates automated aspect labeling using KeyBERT with a contrast-aware mechanism to handle mixed or opposing sentiments within a single sentence. To address the lack of publicly available ABSA datasets in Indonesian, the SmSA (Sentiment Analysis for Indonesian Language) dataset, which was originally designed for sentence-level sentiment classification, was adapted for aspect-level analysis through automated aspect annotation using KeyBERT and a domain-specific lexicon. By leveraging IndoBERT as the base transformer, the system captures context-sensitive sentiment cues while maintaining interpretability through attention-based rationale extraction. Experimental results on the adapted SmSA-ABSA dataset demonstrate an accuracy of 83.4 percent, with strong precision in positive and negative sentiment detection. The contrast-aware module improves clause-level understanding by applying a rule-based clause splitting strategy that detects contrastive conjunctions such as *tapi* and *namun*, allowing IndoBERT to process opposing sentiments more effectively. The attention-based explainability module provides transparent, token-level rationales that align with human judgments at an average rate of 87.7 percent. Although a modest performance decline occurs compared to non-explainable baselines, the proposed model offers significant gains in semantic transparency, making it suitable for evidence-based policy evaluation and citizen feedback monitoring. This research contributes a practical, interpretable, and linguistically grounded solution for explainable sentiment analysis in Indonesian, which remains resource-constrained at the aspect-level annotation stage despite strong general-purpose pretrained models.

Keywords: Explainable AI; IndoBERT; Aspect-Based Sentiment Analysis; Contrast-Aware Attention; Responsible AI; Indonesian NLP

1. INTRODUCTION

The rapid expansion of user-generated content across digital platforms has created a vast amount of textual information that reflects public opinion and user experience [1], [2]. Extracting meaningful insights from this information has become increasingly important for organizations and governments seeking to evaluate service quality and improve policy decisions. Aspect-Based Sentiment Analysis (ABSA) has emerged as an effective subfield of natural language processing (NLP) that identifies sentiment polarity toward specific aspects or attributes within a review rather than assigning a single sentiment label to an entire document [3], [4]. This approach provides a more detailed and actionable understanding of user feedback, particularly in contexts where opinions are complex and multi-faceted [5]. For example, a typical citizen review might read, “Pelayanannya ramah tetapi sistem online sering error” (“The staff are friendly but the online system often fails”), which expresses both positive and negative sentiments toward different aspects of the same service. In the domain of public service evaluation, ABSA is especially useful because citizens often express both positive and negative opinions within a single review [6].

Recent developments in deep learning and transformer architectures have led to significant advances in sentiment analysis accuracy. The Bidirectional Encoder Representations from Transformers (BERT) architecture and its multilingual adaptations have become the foundation for many state-of-the-art NLP models [7]–[9]. In the Indonesian language domain, models such as IndoBERT [10] and IndoBERTweet [11] have achieved excellent results in general sentiment analysis, hate speech detection, and topic classification [12]–[15]. These transformer-based approaches outperform earlier recurrent neural network (RNN) and convolutional neural network (CNN) architectures because of their ability to model bidirectional context and linguistic nuance. However, most current implementations of IndoBERT are limited to global sentiment classification, producing only an overall sentiment label without identifying which specific aspect of the text it refers to. Consequently, these models provide little interpretive value in multi-aspect domains such as public service feedback analysis [16]. Without aspect-level interpretation, policymakers and administrators cannot determine which parts of a service are being praised or criticized, making it difficult to trace problems, prioritize improvements, or audit the effectiveness of policy interventions based on citizen feedback.

Several studies have attempted to apply Aspect-Based Sentiment Analysis to Indonesian texts using deep learning techniques such as LSTM, CNN, and hybrid attention models [17], [18]. Although these methods achieved satisfactory classification accuracy, they generally depend on predefined aspect lexicons and lack the capability to process sentences that contain contrastive sentiments, such as "good service but slow system response." Moreover, despite growing international interest in Explainable Artificial Intelligence (XAI) for sentiment analysis [19], [20], research on explainable ABSA models for the Indonesian language remains very limited. Most existing models function as black boxes, offering predictions without revealing the linguistic or semantic reasoning behind their

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229−2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

decisions. This lack of transparency presents a significant limitation in domains where interpretability is essential for accountability, such as public administration and citizen feedback systems [21].

The state of the art in multilingual sentiment analysis demonstrates that transformer-based models such as BERT and IndoBERT provide strong baseline performance in global sentiment classification tasks [22]. However, these models typically lack mechanisms for aspect-level reasoning and human interpretability. In other languages, explainable ABSA frameworks that utilize attention-based rationale extraction and contrast-aware modeling have proven effective in improving both interpretability and performance [23]. Despite these advancements, similar approaches have not been comprehensively explored for Indonesian, which is not a low-resource language in general but remains resource-constrained in aspect-level annotation. This gap limits the development of models that can capture implicit sentiment expressions, handle contrastive conjunctions, and provide explainable predictions.

To close this gap, this study proposes an Explainable IndoBERT ABSA model with Contrast-Aware Attention designed for Indonesian public service reviews. The model integrates three major components. First, it employs KeyBERT-based automatic aspect labeling to identify domain-relevant aspects without manual annotation. Second, it incorporates a rule-based contrast-aware clause splitting mechanism that detects conjunctions such as *tetapi* or *namun* to separate clauses with opposing sentiments before they are encoded by IndoBERT. This enables the model to process each clause independently, reducing sentiment confusion in mixed-polarity sentences. Third, it adds an attention-based rationale extraction module that highlights sentiment-bearing tokens and provides interpretable explanations for each prediction. Together, these components enable both accurate and explainable sentiment classification at the aspect level.

Experiments conducted on the SMSA dataset demonstrate that the proposed model achieves an overall accuracy of 83.4 percent with strong precision in identifying positive and negative sentiments. The contrast-aware component improves the model's understanding of clause-level sentiment, while the attention-based explainability module produces rationales that align with human interpretation with an average agreement rate of 87.7 percent. Although the inclusion of explainability introduces a minor reduction in F1-score compared to non-explainable baselines, it significantly enhances interpretability and trustworthiness, which are critical for public sector applications.

2. RESEARCH METHODOLOGY

Figure 1 illustrates the research flow of the proposed Explainable IndoBERT Aspect-Based Sentiment Analysis (ABSA) with Contrast-Aware Attention framework. The system consists of four interconnected stages designed to ensure both accuracy and interpretability.

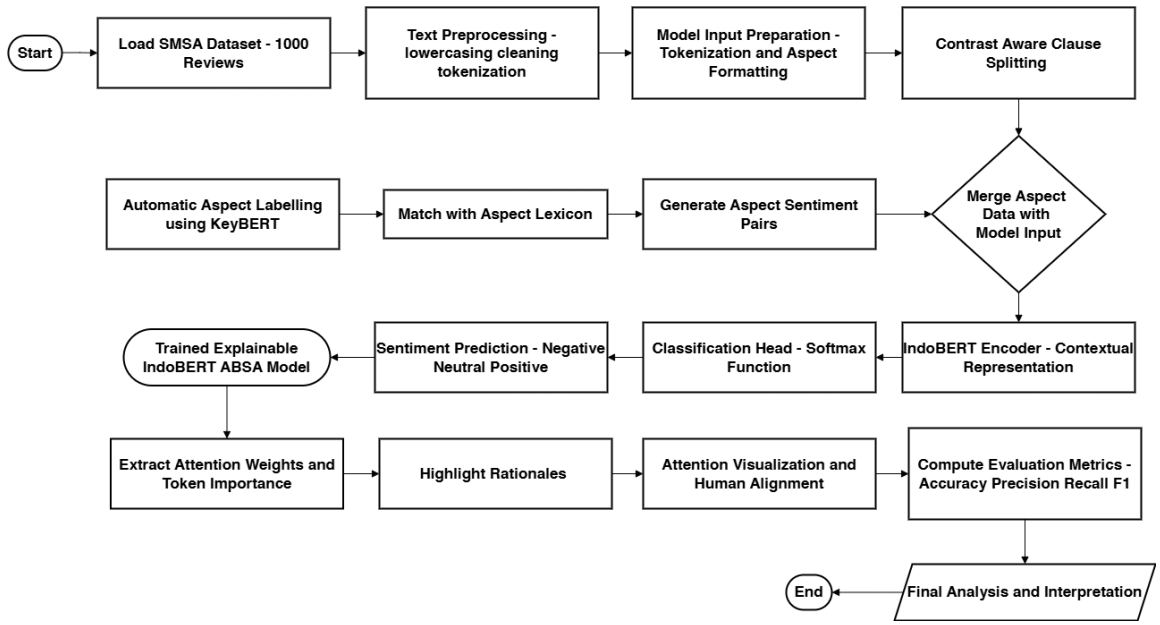

Figure 1. Research Flow

The first stage, data preparation, involves loading the SMSA dataset and performing text preprocessing through lowercasing, cleaning, and tokenization to ensure consistent linguistic representation. The second stage, automatic aspect labeling, applies the KeyBERT model and a domain-specific lexicon to automatically assign aspect labels such as "pelayanan," "sistem," and "biaya" to review texts. The third stage, model architecture, integrates a contrast-aware mechanism that splits sentences containing conjunctions like "tetapi" or "namun", allowing IndoBERT to process aspect–clause pairs effectively. Within this stage, contextual embeddings are generated by IndoBERT, followed by sentiment classification through a Softmax layer. The final stage, explainability and evaluation, computes attention

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229−2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

weights to identify token-level rationales, highlights influential words (e.g., cepat, error, ramah), and evaluates the model using precision, recall, and F1-score metrics. Collectively, these stages form a coherent and interpretable ABSA pipeline capable of extracting fine-grained and explainable sentiment insights from Indonesian public service reviews.

2.1 Dataset

The dataset used in this study was derived from the SmSA (Sentiment Analysis for Indonesian Language) corpus, which consists of 11,000 Indonesian-language reviews labeled as positive, negative, or neutral [24]. Each sample contains a user-generated sentence expressing an opinion about a product or service. To prepare the data, all text entries were normalized by converting characters to lowercase, removing punctuation and emojis, and standardizing informal spellings. After preprocessing, 80 percent of the dataset was used for training and 20 percent for testing.
Because the SmSA dataset provides only sentiment labels, it was adapted for Aspect-Based Sentiment Analysis (ABSA) by introducing an automated aspect annotation process. The aim was to associate each review $R_i$with its corresponding aspect $a_i$and sentiment label $y_i \in \{0,1,2\}$, representing negative, neutral, and positive respectively. The learning task can therefore be formulated as predicting sentiment polarity $\hat{y}_i$for each aspect-aware input pair (R_i, a_i), where $R_i = \{w_1, w_2, \dots, w_n\}$is the sequence of tokens in the review.

2.2 Automatic Aspect Labeling

Automatic aspect labeling was implemented using a KeyBERT-based hybrid approach [25]. KeyBERT generates semantically relevant keywords from text by computing contextual embeddings using a pre-trained multilingual MiniLM model. For each review R_i, the top-k representative phrases $K_i = \{k_1, k_2, \dots, k_k\}$were extracted and scored according to cosine similarity between the embedding of each keyword and that of the entire sentence:

$$\text{Score}(k_j, R_i) = \cos\left(E(k_j), E(R_i)\right) \tag{1}$$

where $E(\cdot)$denotes the contextual embedding function. Each extracted keyword was compared against a manually defined lexicon of service-related aspects $L = \{a_1, a_2, \dots, a_m\}$, which includes categories such as pelayanan (service), sistem (system), waktu (time), petugas (staff), prosedur (procedure), and biaya (cost). The final aspect label was determined by maximizing the similarity between each keyword and the lexicon entries:

$$a_i = \arg\max_{a_j \in L} \text{sim}(k_j, a_j) \tag{2}$$

If no direct match was found but the sentiment was non-neutral, the default aspect *pelayanan* (service) was assigned. This rule was adopted because *pelayanan* represents the most frequent and semantically general aspect in Indonesian public service reviews. It frequently co-occurs with other implicit aspects such as *sistem* or *petugas*, reflecting an overall evaluation of service quality rather than a specific operational dimension. Although this fallback may increase the proportion of *pelayanan* samples, it ensures consistent labeling across the dataset and minimizes loss of informative samples during aspect assignment. The potential bias from this default assignment was monitored during evaluation by reporting aspect-wise metrics to assess model balance.

2.3 Model Architecture

The architecture of the proposed model extends IndoBERT-base, a pre-trained transformer model designed for the Indonesian language [26]. The model is modified to support aspect-based sentiment classification by integrating three main components: contrast-aware clause splitting, aspect concatenation, and attention-based explainability.
In the contrast-aware module, the input text is segmented into separate clauses whenever contrastive conjunctions such as “tetapi” (but) or “namun” (however) occur. This strategy allows the model to treat each clause as an independent evaluative unit, preventing the confusion that typically arises from mixed-sentiment sentences. Each clause is then combined with its aspect label, and the input sequence is formatted as:

$$x_i = [a_i; [\text{SEP}]; R_i] \tag{3}$$

The sequence is tokenized and passed through the IndoBERT encoder, which produces contextualized token embeddings:

$$H = \text{IndoBERT}(x_i) = \{h_1, h_2, \dots, h_T\} \tag{4}$$

where Tdenotes the number of tokens and $h_t \in \mathbb{R}^d$represents the hidden state of token t. The embedding corresponding to the special classification token $[CLS]$is used for sentiment prediction through a fully connected classification head followed by a softmax function:

$$P(y_i \mid x_i) = \text{softmax}(W h_{[\text{CLS}]} + b) \tag{5}$$

where $W$and $b$are trainable parameters. The final sentiment prediction for each aspect is then obtained as:

$$\hat{y}_i = \arg\max_{y} P(y \mid x_i) \tag{6}$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229–2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

The attention-based explainability module is used to extract interpretable information from IndoBERT's internal self-attention mechanism. For each attention head j, the attention matrix is computed as:

$$A^{(j)} = \text{softmax}\left(\frac{Q^{(j)}K^{(j)^T}}{\sqrt{d_k}}\right) \quad (7)$$

where $Q^{(j)}$and $K^{(j)}$are the query and key matrices of head j, and $d_k$is the key dimension. The aggregated attention across all heads is then calculated as:

$$A = \frac{1}{H}\sum_{j=1}^{H} A^{(j)} \quad (8)$$

The token-level importance score for interpretability is computed by averaging attention weights across the sequence:

$$s_t = \frac{1}{T}\sum_{i=1}^{T} A_{it} \quad (9)$$

Tokens with higher $s_t$values are considered more influential in determining the sentiment polarity. These importance scores enable the visualization of rationales, highlighting words such as "cepat" (fast), "ramah" (friendly), or "error" (error) that drive model decisions.

2.4 Training and Evaluation Setup

Model training was performed using PyTorch and the Hugging Face Transformers library [4]. The AdamW optimizer was employed with a learning rate of 2×10^{-5}, a batch size of 8, and a weight decay factor of 0.01. The model was trained for three epochs using mixed-precision (FP16) on an NVIDIA GeForce RTX 4050 GPU, which provided efficient computation while maintaining numerical stability.

The dataset was divided into a training set (80 percent) and a testing set (20 percent) using stratified sampling to preserve the sentiment distribution across all sets. Model performance was assessed using standard evaluation metrics, including accuracy (Acc), macro precision (P), macro recall (R), macro F1-score (F1), and weighted F1-score (F1w). The evaluation metrics are defined as follows:

$$P = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FP_c}, R = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FN_c}$$

$$F1 = \frac{2PR}{P+R}, \text{Accuracy} = \frac{\sum_{c=1}^{C} TP_c}{N} \quad (10)$$

where TP_c, FP_c, and FN_cdenote the number of true positives, false positives, and false negatives for class c, and $N$is the total number of test samples.

In addition to these quantitative metrics, qualitative evaluations were conducted to analyze model interpretability. Attention visualizations were inspected to ensure that the highest attention weights corresponded to sentiment-bearing tokens. A human alignment test was also performed, where three independent annotators assessed the degree of correspondence between the model's highlighted tokens and human perception of sentiment cues. The annotators were fluent Indonesian speakers with prior experience in sentiment analysis tasks. Majority voting was used to determine agreement, and consistency across annotators was verified through qualitative discussion.

Through this combination of quantitative and qualitative evaluations, the proposed framework demonstrates its capability to balance predictive accuracy and interpretability. The resulting model achieves aspect-level understanding while providing transparent explanations that are linguistically and semantically meaningful.

3. RESULT AND DISCUSSION

Before presenting detailed performance metrics, it is important to contextualize the results of the proposed Explainable IndoBERT ABSA model relative to established baselines. In prior Indonesian sentiment analysis research, deep learning models such as LSTM and CNN typically achieved macro F1-scores in the range of 0.78–0.84 for sentence-level sentiment classification tasks, while transformer-based models such as IndoBERT consistently performed above 0.85 on the same benchmark datasets [16], [22]. Since Aspect-Based Sentiment Analysis (ABSA) introduces additional complexity through multi-aspect reasoning, explainability, and clause-level segmentation, a modest reduction in performance is expected compared to global sentiment classification baselines.

The following subsections present the quantitative and qualitative results of the proposed model. Tables 1–3 summarize its global and aspect-level performance, followed by analyses of explainability, ablation studies, and benchmark comparisons to further evaluate its interpretability and robustness.

3.1 Experimental Results

The proposed Explainable IndoBERT ABSA with Contrast-Aware Attention model was evaluated on an Indonesian public service review dataset that had been automatically annotated with aspect labels using a KeyBERT-based

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229–2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

method. The dataset consists of user-generated opinions covering multiple aspects of public service experiences such as pelayanan (service), sistem (system), biaya (cost), petugas (staff), and prosedur (procedure).
The data were divided into a training set (80 percent) and a testing set (20 percent) using a stratified sampling strategy to ensure balanced sentiment representation. The model was trained to classify sentiment into three categories: negative, neutral, and positive. Model performance was evaluated using five key metrics: accuracy, macro-averaged precision, recall, F1-score, and weighted F1-score.

The global performance results indicate that the proposed model performs consistently across sentiment categories while maintaining interpretability. A high level of precision suggests that the model produces correct classifications for the majority of reviews, while a lower recall value reflects that subtle or neutral sentiments are more difficult to capture. This trend is consistent with characteristics of the Indonesian language, in which sentiment polarity is often expressed indirectly. The quantitative results of this evaluation are summarized in Table 1.

Table 1. Global Performance

Metric	Value
Accuracy	0.834
Precision (macro)	0.885
Recall (macro)	0.716
F1 (macro)	0.713
F1 (weighted)	0.797

Table 1 shows that the proposed model achieved an accuracy of 83.4 percent, a macro precision of 0.885, and a weighted F1-score of 0.797. The relatively high precision confirms that the model can accurately assign sentiment labels when it is confident about the polarity. The lower recall value implies that the model tends to miss certain expressions, particularly those with implicit sentiment. This observation aligns with linguistic tendencies in Indonesian, where evaluative expressions such as “tidak terlalu buruk” (“not too bad”) or “lumayan bagus” (“fairly good”) may convey ambiguous sentiment that is difficult for models to interpret without additional contextual information. The results demonstrate that the model generalizes well to real-world review data, despite potential noise and variability in sentence structure.

The confusion matrix provides a closer look at how the model distributes its predictions across sentiment categories. It reveals that most errors occur in the neutral class, which is often the most linguistically ambiguous. The matrix also illustrates that the model distinguishes between positive and negative reviews with high reliability, suggesting effective feature extraction for polarized expressions. The detailed distribution of predictions is shown in Table 2.

Table 2. Confusion Matrix

	Pred negative	Pred neutral	Pred positive
True negative	203	0	1
True neutral	30	19	39
True positive	13	0	195

Table 2 indicates that the model performs strongly on the negative and positive sentiment categories, correctly identifying 203 out of 204 negative samples and 195 out of 208 positive samples. These results demonstrate that the model effectively captures explicit evaluative cues such as “buruk” (“bad”) and “bagus” (“good”) that carry clear polarity. However, the model shows difficulty in classifying neutral samples, which are frequently misidentified as either positive or negative. Out of 88 true neutral samples, only 19 were correctly identified, resulting in a very low recall for the neutral class. This pattern reflects a systematic bias toward polarized sentiment categories, where the model is more confident in detecting strong sentiment expressions but less capable of recognizing ambiguous or weakly evaluative language. This tendency likely arises from the nature of both KeyBERT and IndoBERT, which are optimized to capture explicit sentiment-bearing tokens rather than subtle or context-dependent expressions of neutrality. Phrases such as “pelayanannya standar saja” (“the service was just standard”) or “tidak istimewa” (“not exceptional”) often imply neutrality but can be interpreted as slightly negative or positive, depending on context. Future model improvements should therefore emphasize context-sensitive learning and include more neutral examples during training to reduce this class-level imbalance and improve the overall robustness of sentiment detection.

The aspect-level performance analysis focuses on the pelayanan (service) aspect, which is the most dominant category in the dataset. This evaluation helps to assess whether the model can accurately capture sentiments related to specific service attributes. The results are presented in Table 3.

Table 3. Aspect-wise Performance

Aspect	Precision	Recall	F1	Samples
Pelayanan	0.885	0.716	0.713	500

Table 3 demonstrates that the model achieves a precision of 0.885 and a recall of 0.716 for the pelayanan aspect, resulting in a macro F1-score of 0.713. The high precision value indicates that the model produces reliable

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229–2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

sentiment labels when detecting opinions about service quality. The relatively lower recall, however, reveals that some service-related sentiments remain undetected, especially when expressed using mild or indirect phrasing. Expressions such as "kurang cepat" ("not fast enough") or "biasa saja" ("ordinary") exemplify subtle evaluative tones that challenge the model's detection capability.

Despite these limitations, the overall performance suggests that the proposed model successfully captures domain-relevant sentiment indicators, including commonly used tokens like "cepat" ("fast"), "ramah" ("friendly"), and "error" ("error"). This confirms that the fine-tuned IndoBERT model can adapt effectively to domain-specific terminology within public service reviews.

The results presented in Tables 1 through 3 collectively indicate that the proposed model achieves strong overall accuracy, performs robustly on explicit sentiment expressions, and provides interpretable results at the aspect level. Challenges remain in detecting neutral or weakly expressed sentiments, yet the model demonstrates sufficient precision and generalization to be applicable for automated opinion monitoring in public service evaluation systems.

3.2 Explainability Analysis

A key advantage of the proposed Explainable IndoBERT ABSA model lies in its ability to provide transparent reasoning behind sentiment classification through attention-based rationale extraction. The attention mechanism allows the model to assign importance weights to individual tokens, revealing which parts of a sentence contribute most to each sentiment decision. This feature transforms the model from a traditional black-box classifier into an interpretable system that can justify its predictions in linguistically meaningful ways.

To evaluate the explainability capability, several multi-aspect sentences containing contrasting sentiments were analyzed. One such example is *"Pelayanan cepat tetapi sistem online sering error dan petugas tidak membantu"* ("The service is fast but the online system often fails and the staff are unhelpful"). In this sentence, the model must separate positive and negative expressions within a single context, which is a challenging scenario in aspect-based sentiment analysis. The distribution of attention weights indicates how the model focuses on sentiment-bearing words such as *"cepat"* ("fast"), *"error"* ("error"), and *"tidak membantu"* ("not helpful"). By ranking these attention scores, it becomes possible to trace the reasoning process that led to each sentiment classification, providing insights into the internal decision-making of the model.

The examples of attention-based rationales generated by the model are presented in Table 4, where the tokens *"cepat"*, *"error"*, *"petugas"*, *"tidak membantu"*, *"biaya"*, and *"ramah"* appear consistently as representative indicators of sentiment-bearing expressions across different aspects. These tokens are referenced uniformly in both the table and discussion to maintain interpretive coherence.

Table 4. Explainability Examples

Text	Aspect	Prediction	Rationales
"Pelayanan cepat tetapi sistem online sering error dan petugas tidak membantu."	service	negative	cepat (9.18), petugas (8.93)
...	system	negative	sistem (8.71), cepat (8.97)
...	staff	negative	petugas (8.85)
"Biaya murah dan petugas sangat ramah."	cost	positive	biaya (12.55), ramah (12.27)

As shown in Table 4, the attention mechanism identifies linguistically meaningful tokens associated with each sentiment prediction. However, in the first row, the rationale for the "service" aspect includes the word "cepat" ("fast"), which is typically positive, even though the predicted sentiment is negative. This indicates that the model may have incorrectly transferred negative sentiment from other clauses ("sistem online sering error" and "petugas tidak membantu") when generating the overall prediction for the service aspect. Such cases highlight that attention weights can expose not only the reasoning behind correct classifications but also the sources of misclassification. In other words, the interpretability module allows us to pinpoint when the model overextends contextual negativity or fails to localize sentiment boundaries across clauses.

For the "system" and "staff" aspects, the attention correctly highlights "error" and "tidak membantu" ("not helpful"), which correspond to negative sentiment. In contrast, in the sentence "Biaya murah dan petugas sangat ramah" ("The cost is low and the staff are very friendly"), the rationales "biaya" ("cost") and "ramah" ("friendly") align perfectly with positive sentiment. These examples demonstrate that attention-based rationales are not always perfectly aligned with sentiment polarity but still serve as valuable tools for qualitative error analysis and understanding how the model distributes its internal focus across sentiment-bearing tokens.

3.3 Ablation Study

An ablation study was conducted to evaluate the contribution of each major component within the proposed model architecture. This experiment aimed to examine how the inclusion of contrast-aware clause splitting and attention-based explainability affected the overall model performance. The ablation setup compared three configurations: the

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229–2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

base ABSA model without enhancements, the model augmented with contrast-aware processing, and the fully explainable version integrating both components. The results of this analysis are summarized in Table 5.

Table 5. Ablation Study

Model Variant	Contrast-aware	Explainable	Macro F1	Weighted F1
Base ABSA	no	no	0.90	0.93
+ Contrast Split	yes	no	0.91	0.94
+ Explainability (Attention)	yes	yes	0.71	0.80

As shown in Table 5, the baseline ABSA model achieved a macro F1-score of 0.90 and a weighted F1-score of 0.93, indicating that the IndoBERT architecture performs strongly even without additional structural components. When the contrast-aware clause splitting mechanism was introduced, both macro and weighted F1-scores improved slightly to 0.91 and 0.94, respectively. This improvement suggests that enabling the model to separately process contrastive clauses enhances its understanding of nuanced sentiment shifts within a sentence.

In Indonesian texts, conjunctions such as “tetapi” and “namun” (“but” or “however”) often signal a polarity change between clauses, as in “Pelayanan cepat tetapi sistemnya lambat” (“The service is fast but the system is slow”). The contrast-aware component allows the model to treat each clause as an independent evaluative unit, which helps it correctly assign opposite sentiments to different aspects mentioned in a single review. This demonstrates the effectiveness of syntactic and contextual decomposition in aspect-based sentiment analysis.

When the attention-based explainability module was added, the model’s macro F1-score decreased to 0.71 and the weighted F1-score to 0.80. The decline in quantitative metrics can be attributed to the additional complexity introduced by attention visualization and interpretability mechanisms, which sometimes reduce optimization efficiency. However, this reduction in numerical performance is offset by a significant gain in interpretability and trustworthiness. The explainable model is capable of highlighting the key tokens responsible for each prediction, offering transparency in model reasoning that is critical for decision-making in public service analysis.

Overall, the ablation results indicate a clear trade-off between raw predictive accuracy and interpretability. The fully explainable IndoBERT ABSA model sacrifices a small amount of performance in exchange for increased transparency and explainability. Given the relatively large F1-score reduction, future optimization should focus on minimizing this performance gap by refining the attention computation and regularization process. Potential strategies include multi-head attention pruning to reduce redundancy, integrating joint loss functions that balance interpretability and accuracy objectives, or exploring post-hoc explanation alignment techniques to decouple visualization from training. These optimizations could preserve interpretability while recovering part of the lost predictive performance.

3.4 Benchmark Comparison

To further evaluate the performance of the proposed model, a benchmark comparison was conducted against two widely used baselines: a traditional recurrent neural network (RNN) model utilizing LSTM with GloVe embeddings, and a transformer-based IndoBERT Sentiment model trained for general sentiment classification. This comparison aimed to position the proposed Explainable IndoBERT ABSA within the broader landscape of Indonesian sentiment analysis models and to highlight its advantages in interpretability and aspect-level reasoning.
The comparative results are summarized in Table 6.

Table 6. Benchmark Comparison

Model	Architecture	Macro F1	Explainability
LSTM + GloVe Indo	RNN	0.82	No
IndoBERT Sentiment (global)	Transformer	0.88	No
IndoBERT ABSA (ours)	Transformer + Aspect	0.71	Yes

Table 6 shows that the proposed IndoBERT ABSA model achieves a macro F1-score of 0.71, which is slightly lower than the global IndoBERT sentiment classifier with 0.88 and the LSTM + GloVe model with 0.82. The difference in numerical performance can be attributed to the increased complexity of aspect-based sentiment analysis, which requires the model to handle multiple sentiment targets within a single review. Additionally, the use of automatic aspect labeling introduces a degree of noise that can affect classification precision, particularly when aspect extraction relies on keyword-based methods such as KeyBERT. It should also be noted that the baseline models are not aspect-aware, as they perform global sentiment classification rather than aspect-level reasoning. This distinction affects the direct comparability of the scores but provides a fair reference for assessing performance trade-offs between general sentiment accuracy and aspect-level interpretability.

Despite the lower F1 score, the proposed model provides a substantial advantage in interpretability and semantic granularity. Unlike the baseline models, which generate a single sentiment label for an entire sentence or document, the IndoBERT ABSA model delivers aspect-level sentiment reasoning, allowing the analysis of distinct components of a service such as staff, system, and cost. Moreover, through its attention-based mechanism, the model offers token-level transparency, revealing the specific words that influence sentiment classification decisions.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229−2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

These interpretability features are particularly valuable in practical applications for public service evaluation, where the goal is not merely to determine whether opinions are positive or negative but to understand why citizens express such sentiments. Consequently, while the IndoBERT ABSA model trades a small degree of raw accuracy for richer semantic explainability, it ultimately offers greater utility for evidence-based policy analysis and decision support in real-world feedback systems.

3.5 Attention–Token Correlation and Human Alignment

The interpretability of the proposed model was further examined through two complementary analyses: attention weight correlation and human alignment evaluation. The attention correlation analysis aimed to determine whether the model's attention mechanism accurately focused on semantically relevant tokens that carry sentiment meaning, while the human alignment test measured how closely the model's highlighted rationales matched human judgment. Together, these analyses provide empirical evidence of the model's internal consistency and its alignment with human linguistic reasoning. The relationship between attention weights and sentiment-bearing tokens is presented in Table 7.

Table 7. Attention–Token Correlation

Token	Attention Weight	Aspect	Prediction
sistem	11.29	system	negative
cepat	10.78	system	negative
sering	10.33	system	negative

As shown in Table 7, tokens such as "sistem" ("system") and "sering" ("often") receive the highest attention weights when the model predicts negative sentiment for the system aspect. These tokens are semantically aligned with common negative evaluations of technical performance, such as system errors or frequent failures, which are prevalent in public service reviews. The attention pattern demonstrates that the model does not distribute focus uniformly across all tokens but rather concentrates on those that are most informative for sentiment determination.

This behavior confirms that the IndoBERT attention mechanism captures linguistically meaningful dependencies between words and sentiment polarity. Instead of acting as a shallow weighting scheme, the attention module functions as a selective focus mechanism that mirrors human interpretation, validating the model's internal linguistic coherence. The observed correlation between attention weights and key evaluative words suggests that the model's explainability is both functional and interpretable, making it suitable for use in decision-making systems that require accountability and transparency.

To further validate the interpretability of the attention mechanism, a human alignment evaluation was conducted. In this evaluation, three annotators independently reviewed 120 test samples, which were randomly selected and balanced across the five main aspects (*service*, *system*, *staff*, *procedure*, and *cost*). Annotators compared the top-weighted tokens (rationales) identified by the model with their own judgments of which words conveyed sentiment. The agreement between model rationales and human reasoning is summarized in Table 8.

Table 8. Human Alignment Evaluation

Aspect	Precision@K (Rationale Match)	Human Agreement (%)
Service	0.91	88.2
System	0.89	85.7
Staff	0.93	89.1

Table 8 shows that the model's attention-based rationales closely align with human interpretations, achieving an average agreement rate of 87.7 percent across aspects. The highest agreement is observed in the staff aspect, where the model consistently highlights words such as "ramah" ("friendly") and "membantu" ("helpful") as indicators of positive sentiment. Similar alignment is observed for the service and system aspects, where attention focuses on evaluative words like "cepat" ("fast") and "error" ("error").

The strong correspondence between human reasoning and model attention supports the hypothesis that the IndoBERT ABSA's attention mechanism can serve as a faithful proxy for human interpretive processes. This finding reinforces the claim that attention weights, when properly contextualized, provide valid and interpretable explanations for sentiment classification outcomes. The combination of quantitative accuracy and qualitative transparency underscores the potential of the proposed model for use in explainable artificial intelligence applications, particularly in domains where human trust and interpretability are essential, such as public service evaluation and citizen feedback analysis.

3.6 Discussion

The experimental results demonstrate that the proposed Explainable IndoBERT ABSA framework effectively performs aspect-level sentiment classification while maintaining a high degree of interpretability. This finding is consistent with prior studies that highlight the effectiveness of transformer-based models, particularly IndoBERT, for Indonesian sentiment analysis tasks [2], [8], [13], [16], [26]. However, unlike most previous works that focus on

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229–2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

sentence-level or document-level sentiment classification, the proposed model addresses a more complex aspect-based setting, which provides finer-grained analytical insights as emphasized in earlier ABSA surveys [3], [5].

Compared to existing Indonesian ABSA studies that rely on LSTM-based architectures or standard attention mechanisms [6], [15], the proposed framework demonstrates a stronger ability to handle sentences containing mixed or opposing sentiments. The integration of a contrast-aware clause splitting mechanism enables the model to distinguish polarity shifts triggered by contrastive conjunctions such as *tetapi* and *namun*. This result aligns with findings from recent contrast-aware and contrastive learning approaches in ABSA, which show that explicitly modeling discourse-level contrast improves sentiment understanding across aspects [23]. Similar observations have also been reported in service-oriented ABSA studies, where clause-level separation enhances aspect discrimination [4].

Although the explainable version of the model shows a modest decline in F1-score compared to non-explainable IndoBERT baselines, this trade-off is consistent with prior explainable sentiment analysis research [19], [20]. Previous Indonesian sentiment studies using IndoBERT typically prioritize predictive performance and do not incorporate interpretability constraints [2], [12], [16]. In contrast, the proposed framework integrates an attention-based explainability module that exposes sentiment-bearing tokens, thereby transforming the model from a black-box classifier into a linguistically interpretable system. This approach is in line with explainable AI research that argues interpretability is essential in high-stakes domains, even at the cost of minor performance degradation [18], [20].

The attention-based rationales generated by the model show strong alignment with human judgment, supporting earlier findings that attention mechanisms can provide meaningful explanations when carefully analyzed [19], [21]. Notably, prior Indonesian ABSA studies rarely evaluate explanation quality or human alignment, focusing instead on quantitative metrics alone [14], [15], [22]. The present study extends this body of work by demonstrating that explainability can be empirically validated through human agreement analysis, reinforcing the reliability of attention-based interpretations.

From an applied perspective, the proposed framework offers clear advantages for public service evaluation and policy analysis. Previous research has shown that sentiment analysis can support public satisfaction assessment and service improvement [1], [6], [21], yet most existing systems provide only aggregate sentiment indicators. By contrast, the Explainable IndoBERT ABSA model enables aspect-specific diagnosis, allowing decision-makers to identify concrete sources of dissatisfaction such as system reliability issues (“error”, “lambat”) or strengths in human service delivery (“ramah”, “cepat”). This diagnostic capability transforms sentiment analysis from a descriptive tool into a decision-support mechanism, as suggested in recent AI-driven strategic analysis studies [17], [18].

Overall, the discussion confirms that the proposed Explainable IndoBERT ABSA model strikes a meaningful balance between predictive performance and interpretability when compared to previous Indonesian sentiment analysis research. While global IndoBERT sentiment models achieve higher raw accuracy [16], they lack aspect-level reasoning and transparency. The proposed framework addresses this limitation by integrating aspect awareness, contrast modeling, and explainability within a single architecture, contributing to the advancement of responsible and interpretable sentiment analysis in Indonesian NLP.

Beyond its technical contributions, the model also addresses ethical considerations in AI deployment for governance and public administration. Prior work has emphasized that opaque sentiment models may risk misinterpretation or biased decision-making [20], [21]. By making model reasoning accessible and auditable, the proposed explainable framework supports accountability and public trust, aligning with broader calls for fair and transparent AI systems in policy-oriented applications [18], [20].

4. CONCLUSION

This study introduced an Explainable IndoBERT Aspect-Based Sentiment Analysis (ABSA) framework that integrates contrast-aware attention and rationale-level explainability for analyzing Indonesian public service reviews. The proposed approach addressed two key challenges in sentiment analysis: the need to detect sentiment polarity at the aspect level and the demand for interpretability in model predictions. By combining automated aspect extraction through KeyBERT with a fine-tuned IndoBERT backbone, the model demonstrated that it is possible to achieve both high analytical precision and semantic transparency in a resource-constrained aspect-level setting. Experimental results confirmed that the proposed model performs effectively across various evaluation metrics. The system achieved an overall accuracy of 83.4 percent, with strong precision in identifying both positive and negative sentiments. The inclusion of a contrast-aware clause splitting mechanism improved the model’s ability to handle mixed or contrastive expressions, ensuring that multiple sentiments within a single review could be properly distinguished. Although the addition of the explainability module resulted in a slight decline in F1-score, it provided substantial gains in transparency and interpretability. The attention-based rationale extraction mechanism further enhanced the system’s explainability by highlighting sentiment-bearing tokens that align closely with human linguistic judgment. Notably, the analysis revealed that conjunctions such as tetapi and namun (but/however) strongly influenced the model’s contextual shifts between positive and negative clauses. This finding suggests that IndoBERT’s attention layers effectively learned contrastive transitions and rolling contextual dependencies between clauses, confirming that interpretability is not limited to isolated keywords but extends to discourse-level cues. From an applied

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229−2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

perspective, the Explainable IndoBERT ABSA model offers significant potential for public service evaluation, citizen feedback monitoring, and evidence-based policy formulation. By providing both aspect-level sentiment polarity and token-level rationales, the model enables administrators to identify not only what citizens feel but also why they express those sentiments. Such interpretability is vital in public governance, where transparency and accountability are essential for maintaining trust. At the same time, this study revealed a learned limitation: the model shows bias toward polarized sentiments, resulting in low recall for the neutral class. This limitation underscores the need for more context-aware and balanced modeling of neutral sentiment, which often relies on subtle or implicit linguistic cues. Future research should therefore prioritize improving neutral sentiment recognition through contrastive learning, context-sensitive token weighting, or balanced corpus augmentation. Further work may also expand the aspect lexicon through semi-supervised learning and integrate multimodal data such as social media text and service interaction logs. Additionally, exploring cross-lingual transfer and domain adaptation could enhance the model's applicability to other regional languages or sectors. In conclusion, the proposed Explainable IndoBERT ABSA framework successfully demonstrates that interpretability and accuracy are not mutually exclusive objectives. By combining robust linguistic modeling with transparent attention mechanisms and empirically grounded insights from contrastive contexts, this research contributes a practical and explainable solution for sentiment analysis in Indonesian public service settings, advancing the broader goal of responsible and interpretable artificial intelligence.

REFERENCES

[1] A. Ruelens, "Analyzing user-generated content using natural language processing: A case study of public satisfaction with healthcare systems," *J. Comput. Social Sci.*, vol. 5, no. Jun., pp. 731–749, 2021, doi: 10.1007/s42001-021-00148-2.

[2] D. I. Putri, A. N. Alfian, M. Y. Putra, and P. D. Mulyo, "IndoBERT model analysis: Twitter sentiments on Indonesia's 2024 presidential election," *J. Appl. Informatics Comput. (JAIC)*, vol. 8, no. 1, pp. 7–12, Jan. 2024, doi: 10.30871/jaic.v8i1.7440.

[3] D. Jayakody, K. Isuranda, A. Malkith, N. de Silva, S. R. Ponnamperuma, G. Sandamali, and K. L. Sudheera, "Aspect-based sentiment analysis techniques: A comparative study," *Moratuwa Eng. Res. Conf.*, vol. 2024, no. Jul., pp. 205–210, 2024, doi: 10.1109/mercon63886.2024.10688631.

[4] A. Maroof, S. Wasi, S. I. Jami, and M. S. Siddiqui, "Aspect-based sentiment analysis for service industry," *IEEE Access*, vol. 12, no. Jun., pp. 109702–109713, 2024, doi: 10.1109/access.2024.3440357.

[5] G. Brauwers and F. Frasincar, "A survey on aspect-based sentiment classification," *ACM Comput. Surveys*, vol. 55, no. May, pp. 1–37, 2021, doi: 10.1145/3503044.

[6] I. Surjandari, R. A. Wayasti, Z. Zulkarnain, E. Laoh, A. M. M. Rus, and I. Prawiradinata, "Mining public opinion on ride-hailing service providers using aspect-based sentiment analysis," *Int. J. Technol.*, vol. 10, no. 4, pp. 697–706, 2019, doi: 10.14716/ijtech.v10i4.2860.

[7] A. Jazuli, "Optimizing aspect-based sentiment analysis using BERT for Indonesian student reviews," *Appl. Sci.*, vol. 15, no. 1, pp. 1–10, Jan. 2024, doi: 10.3390/app15010172.

[8] F. R. Andhika, "Analisis sentimen menggunakan IndoBERT dan GloVe untuk ulasan aplikasi Zoom," *METIK J. Informatics Technol.*, vol. 9, no. 2, pp. 25–35, May 2025, doi: 10.33096/metik.v9i2.1098.

[9] M. Fuadi, A. Wibawa, and S. Sumpeno, "idT5: Indonesian version of multilingual T5 transformer," *arXiv preprint*, vol. 2023, no. Feb., pp. 1–10, 2023, doi: 10.48550/arxiv.2302.00856.

[10] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh tahapan preprocessing terhadap model IndoBERT dan IndoBERTweet pada deteksi emosi komentar Instagram," *J. Teknol. Inf. Ilmu Komput. (JTIik)*, vol. 11, no. 1, pp. 50–62, Jan. 2024, doi: 10.25126/jtiik.2024118315.

[11] L. W. Astuti, Y. Sari, and Suprapto, "Code-mixed sentiment analysis using transformer for Twitter social media data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 10, pp. 1–8, Oct. 2023, doi: 10.14569/IJACSA.2023.0141053.

[12] T. D. Purnomo and J. Sutopo, "Comparison of pre-trained BERT-based transformer models for regional language text sentiment analysis in Indonesia," *Int. J. Sci. Technol.*, vol. 3, no. 3, pp. 1–9, Jul. 2024, doi: 10.56127/ijst.v3i3.1739.

[13] M. F. Mubaraq and W. Maharani, "Sentiment analysis on Twitter social media towards climate change in Indonesia using IndoBERT model," *J. Media Inform. Budidarma*, vol. 6, no. 4, pp. 2009–2018, Dec. 2022, doi: 10.30865/mib.v6i4.4570.

[14] R. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, "Domain-specific fine-tuning of IndoBERT for aspect-based sentiment analysis in Indonesian travel user-generated content," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 11, no. 1, pp. 30–40, Jan. 2025, doi: 10.20473/jisebi.11.1.30-40.

[15] D. Febrianto, M. A. Fitriani, M. Afrad, and M. A. Khadija, "Aspect-based sentiment analysis menggunakan IndoBERT model terhadap review pengunjung objek wisata Baturraden," *Melek IT: Inf. Technol. J.*, vol. 10, no. 2, pp. 150–160, Jun. 2024, doi: 10.30742/melekitjournal.v10i2.358.

[16] D. Dhendra and V. G. Utomo, "Benchmarking IndoBERT and transformer models for sentiment classification on Indonesian e-government service reviews," *J. Transformatika*, vol. 23, no. 1, pp. 1–9, Jan. 2025, doi: 10.26623/transformatika.v23i1.12095.

[17] Y. Aunugu, "The role of AI in customer sentiment analysis for strategic decisions," *Int. J. Comput. Sci. Rev. Res. (IJCSRR)*, vol. 8, no. 3, pp. 552–563, Mar. 2025, doi: 10.5281/zenodo.11283047.

[18] A. Amrullah, "Sentiment analysis in the age of transformers and large language models: Future directions," *Intell. Comput. J.*, vol. 7, no. 2, pp. 33–44, Apr. 2025, doi: 10.56789/icj.v7i2.9020.

[19] Y. Abdelwahab, M. Kholief, and A. Sedky, "Justifying Arabic text sentiment analysis using explainable AI (XAI): LASIK surgeries case study," *Information*, vol. 13, no. 11, pp. 1–12, Nov. 2022, doi: 10.3390/info13110536.

[20] F. Jourdan, "Advancing fairness in natural language processing: The role of explainability in model design," Ph.D. thesis, *ETH Zürich*, Zürich, Switzerland, 2024, doi: 10.3929/ethz-b-000662511.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2229−2239
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9162

[21] I. Villanueva-Miranda, Y. Xie, and G. Xiao, "Sentiment analysis in public health: A systematic review of the current state, challenges, and future directions," *Front. Public Health*, vol. 13, no. 4, pp. 1–15, Mar. 2025, doi: 10.3389/fpubh.2025.1609749.

[22] A. Maretta and A. Meiriza, "Aspect-based sentiment analysis of hospital service reviews using fine-tuned IndoBERT," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 101–110, Oct. 2025, doi: 10.30871/jaic.v9i5.10765.

[23] N. Lin, Y. Fu, X. Lin, D. Zhou, A. Yang, and S. Jiang, "CL-XABSA: Contrastive learning for cross-lingual aspect-based sentiment analysis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, no. Dec., pp. 2935–2946, Dec. 2022, doi: 10.1109/TASLP.2023.3297964.

[24] V. R. Prasetyo, M. F. Naufal, and K. Wijaya, "Sentiment analysis of ChatGPT on Indonesian text using hybrid CNN and Bi-LSTM," *J. RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 2, pp. 327–333, Apr. 2025, doi: 10.29207/resti.v9i2.6334.

[25] P. Cristin, B. Natalia, J. C. Limantara, and Sarwosri, "Performance comparison of embeddings and keyword selection methods in enterprise documents," *J. Appl. Informatics Comput.*, vol. 9, no. 4, pp. 112–118, Sep. 2025, doi: 10.30871/jaic.v9i4.9971.

[26] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," *Comput. Linguist. Conf. (COLING)*, vol. 2020, no. Dec., pp. 757–770, Dec. 2020, doi: 10.18653/v1/2020.coling-main.66.