Scientific Journal of Informatics
Vol. 12, No. 4, November 2025
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Review: A Hybrid Approach of Aspect-Based Sentiment Analysis and Knowledge Extraction for Evaluating Security Perceptions in Digital Payment Applications

Aisyah Fatihaturrahmah[1], Ken Ditha Tania[2*]
[1,2] Department of Information System, Universitas Sriwijaya, Indonesia

Abstract.
Purpose: The rapid expansion of digital wallets in Indonesia has heightened concerns regarding user security and trust. This study evaluates user sentiment toward the security features of the DANA digital payment application using Aspect Sentiment Classification (ASC), a subtask of Aspect-Based Sentiment Analysis (ABSA). It aims to compare multiple classification models and generate structured, machine-readable sentiment outputs to support knowledge extraction and system integration.
Methods: A total of 4,846 security-related reviews were collected from the Google Play Store using keyword-based filtering, supplemented by 3,000 unfiltered reviews for robustness evaluation. Sentiment labeling was performed using a hybrid rule-based and manual annotation approach. From 300 proportionally sampled reviews (150 positive and 150 negative), the validation achieved 0.8504 accuracy and a Cohen's κ of 0.951, indicating near-perfect agreement. Five models Support Vector Machine (SVM), Random Forest (RF), Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and IndoBERT were evaluated using 5-fold stratified cross-validation with random oversampling to address class imbalance.
Result: IndoBERT achieved the highest performance with 98% accuracy, an F1-score of 0.974, and an AUC-ROC of 0.996, followed by CNN and BiLSTM. Robustness testing across temporal (DANA June–October) and cross-domain (GoPay) datasets confirmed IndoBERT's strong generalization with minimal F1-score variation.
Novelty: Unlike previous ABSA studies that addressed multiple aspects, this research focuses exclusively on the security aspect, providing fine-grained insights into user trust. The integration of XML-based structured output enhances interpretability and interoperability in digital financial sentiment analysis, contributing to the development of more secure and transparent fintech ecosystems.

Keywords: Aspect-based sentiment analysis (ABSA), Security, Sentiment classification, Machine learning, Deep learning, Knowledge extraction
Received July 2025 / **Revised** October 2025 / **Accepted** November 2025

INTRODUCTION

The rapid development of digital technology has driven the widespread adoption of digital wallets, such as DANA, in Indonesia, with DANA emerging as one of the leading platforms offering fast, accessible, and efficient financial transaction solutions [1]. DANA was selected as the focus of this study because it represents one of the largest digital payment ecosystems in the country, with over 135 million active users and integration with the national QRIS infrastructure [2]. Its extensive use in peer-to-peer transfers, bill payments, and online or offline retail transactions underscores its central role in Indonesia's digital economy.

DANA's emphasis on user authentication through e-KTP verification and one-time password (OTP) mechanisms also makes it an appropriate case for examining user sentiment toward digital security. The abundance of user-generated reviews on platforms such as the Google Play Store provides valuable insights into user perceptions, trust, and security-related concerns [3].

Amid growing competition in the digital finance sector, service quality has become a strategic factor in building user trust and loyalty [1], [3]. At the same time, the proliferation of social media and online review platforms has enabled users to share their experiences with financial services, providing valuable datasets for analyzing public sentiment [4]. Reviews often mention specific aspects such as usability, transaction speed, fees, and security which significantly influence user satisfaction [5].

*Corresponding author.
Email addresses: kenya.tania@gmail.com (Tania), aisyahfatiha22@gmail.com (Fatihaturrahmah)
DOI: 10.15294/sji.v12i4.31557

Conventional sentiment analysis typically classifies text as positive, negative, or neutral, but fails to capture which specific aspect is being evaluated [6]. Aspect-Based Sentiment Analysis (ABSA) addresses this limitation by determining sentiment polarity for distinct aspects mentioned in user feedback [7]. Within ABSA, the Aspect Sentiment Classification (ASC) subtask focuses on identifying the sentiment orientation toward predefined aspects [8]. However, many ABSA studies still rely on traditional classifiers such as Naïve Bayes and K-Nearest Neighbor, with limited exploration of advanced deep learning methods that could yield better performance in digital financial contexts [9]. Furthermore, existing ABSA research often uses English-language datasets, limiting applicability to Indonesian-language platforms such as DANA [10].

Recent advances in Indonesian language models, particularly IndoBERT, have shown strong performance in capturing contextual sentiment patterns [11]. Security, as a critical dimension of digital finance, directly impacts user trust. Structuring ABSA results in XML format further enhances interoperability and enables integration into intelligent systems such as monitoring dashboards or recommendation modules [12]. Although prior studies have applied ABSA in financial domains [13], few have produced structured, machine-readable outputs in standardized XML formats, as demonstrated in the FiGAS approach [14].

To address these gaps, this study makes three key contributions. First, it applies the ASC framework to analyze user sentiment toward DANA's security features, offering fine-grained insights into this critical aspect. Second, this study compares the performance of five widely used ABSA classification models: Support Vector Machine (SVM), Convolutional Neural Network (CNN), IndoBERT, Bidirectional Long Short-Term Memory (BiLSTM), and Random Forest (RF) representing both traditional and deep learning paradigms. SVM and RF serve as robust baselines for high-dimensional textual data [15], [16], CNN and BiLSTM effectively capture local and sequential sentiment patterns [17], [18], while IndoBERT provides strong contextual understanding specific to the Indonesian language [19]. This combination enables a comprehensive evaluation of model robustness in aspect-based sentiment classification.

These models were chosen because they represent complementary paradigms in sentiment classification: traditional machine learning (SVM, RF), deep learning (CNN, BiLSTM), and transformer-based architectures (IndoBERT). SVM and RF provide strong baselines for text classification and handle high-dimensional features effectively [15], [16], CNN and BiLSTM capture local and sequential sentiment patterns from unstructured text [17], [18], while IndoBERT offers contextual understanding specific to the Indonesian language [19]. This combination enables a comprehensive comparison of model performance in identifying user sentiment related to digital wallet security.

Third, sentiment outputs are represented in XML format to support structured knowledge extraction and interoperability with intelligent systems, consistent with hierarchical ABSA frameworks [13] and XML-based explainable modeling [20]. To the best of the authors' knowledge, limited studies have examined user sentiment toward the *security* aspect of Indonesian digital wallets using ABSA with comparative model evaluation and XML-based knowledge extraction. By focusing exclusively on DANA's security features, this study provides a deeper understanding of user trust and risk awareness while achieving high annotation consistency (accuracy = 0.8504; κ = 0.951). This single-aspect focus ensures analytical precision and supports reliable model comparison across architectures.

METHODS

This study analyzes user sentiment regarding the security features of the DANA digital wallet application using Aspect Sentiment Classification (ASC), a subtask of Aspect-Based Sentiment Analysis (ABSA). ASC is responsible for determining the sentiment polarity associated with a specific aspect term within a sentence [7], enabling more fine-grained opinion mining that captures user attitudes toward distinct service components [21]. This approach is particularly useful when user concerns focus on a single critical aspect, such as security [22].

Unlike full ABSA pipelines that include both aspect extraction and classification, this study focuses solely on Aspect Sentiment Classification because the aspect of interest DANA Security is predefined based on the study's scope. This focused approach allows deeper model evaluation and more reliable comparison across algorithms [9], [21].

The methodological framework of this study comprises several stages: data collection, data preprocessing, aspect annotation, model implementation, performance evaluation, sentiment output in XML format, and

knowledge extraction. These stages are elaborated in the subsequent section, and the overall workflow is illustrated in Figure 1, which presents the research methodology flowchart.

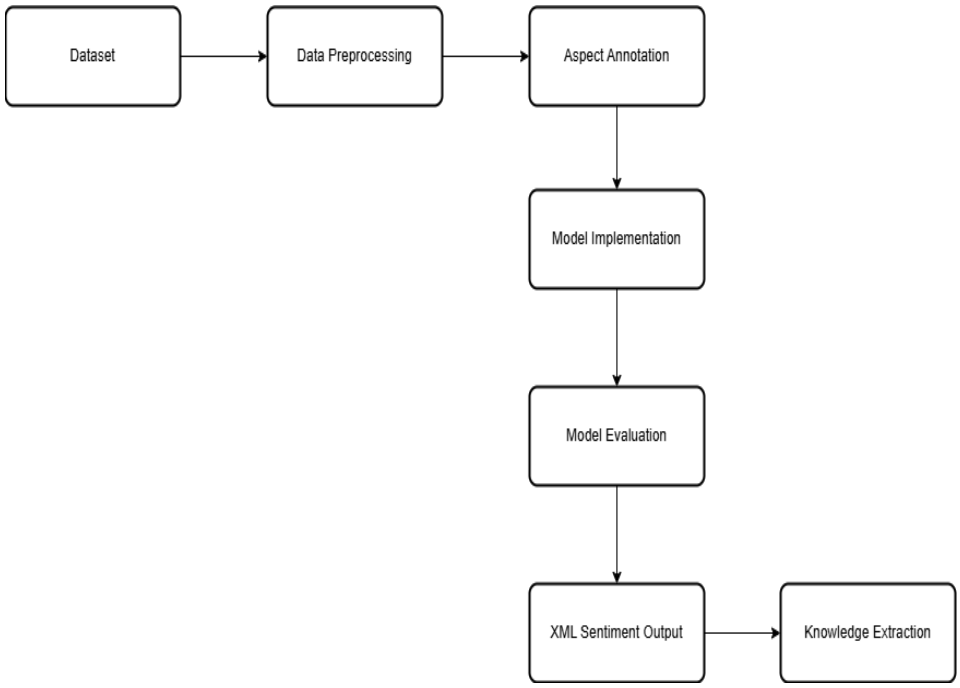

Figure 1. Flowchart Process

Dataset

This study used user reviews of the DANA application collected from the Google Play Store via the *google-play-scraper* Python library. From around 50,000 initial reviews, 4,846 were selected through keyword-based filtering focusing on terms such as *account*, *password*, *verification*, *OTP*, *blocked*, and *security*. This filtering approach follows the standard keyword-based sampling methods commonly applied in previous e-wallet sentiment analysis studies [23].

Sentiment labeling was performed using a rule-based approach, classifying reviews as positive (e.g., "mudah di gunakan aman dan terpercaya") or negative (e.g., "saya kecewa karna aplikasi dana sekarang sudah rawan di hack") based on predefined lexical indicators such as "aman" or "tidak aman".

To ensure reliability and minimize labeling bias, 300 reviews were manually re-annotated by two independent annotators, consisting of 150 positive and 150 negative samples. The comparison between rule-based and manual annotations was conducted to verify the consistency of the automated labeling approach, yielding an accuracy of 0.8504 and an F1-score of 0.8417. The inter-annotator agreement achieved Cohen's $\kappa = 0.951$ (97.58% raw agreement), indicating almost perfect consistency according to established annotation reliability standards [24].

Two supplementary datasets were also employed for robustness evaluation: (1) 3,000 non-filtered DANA reviews collected in October 2025 to assess temporal robustness, and (2) 3,000 GoPay reviews for cross-domain robustness testing. Both datasets underwent identical preprocessing and labeling procedures. To handle class imbalance, random oversampling was applied to the minority class within each cross-validation fold. This dataset preparation aligns with prior Indonesian sentiment studies that applied lexical pattern detection and rule-based implicit aspect identification [24]. Complete dataset statistics are presented in Table 1.

Table 1. Security Review Dataset Statistics

Information	Value
Total reviews after filtering (DANA)	4,846
Additional non-filtered DANA Reviews (October 2025)	3,000
Cross-domain GoPay Reviews	3,000
Number of positive reviews (Main Dataset)	1,244
Number of negative reviews (Main Dataset)	3,602
Manual Annotation samole size	300
Annotation Agreement (Cohen's K)	0.951
Average review length (in words)	21.07

Data Preprocessing

Data preprocessing is essential in sentiment analysis to ensure that textual inputs are clean and model-ready. In this study, user reviews were normalized through several steps: case folding, removal of non-alphabetic characters (e.g., punctuation, emojis, digits), character normalization to handle exaggerated expressions, tokenization, and whitespace cleaning [25].

These processes were implemented in Python using the re library for pattern filtering and pandas for data handling and pandas for data manipulation. The cleaned texts were stored in a clean_content, which served as input for all models: SVM, Random Forest, CNN, BiLSTM and IndoBERT.

For the traditional and deep learning models (SVM, Random Forest, CNN, BiLSTM), text sequences were transformed into numerical representations using TF-IDF vectorization and token indexing, followed by padding to a fixed sequence length to ensure uniform input dimensions. In contrast, IndoBERT employed subword tokenization via the AutoTokenizer from the Hugging Face transformers library, which preserved semantic consistency and contextual relationships within Indonesian-language text.

This comprehensive preprocessing pipeline—encompassing cleaning, normalization, tokenization, and vectorization—follows established best practices in sentiment analysis and has been shown to significantly enhance model stability and performance across various NLP applications [26].

Aspect Annotation

This study applies Aspect-Based Sentiment Analysis (ABSA) using a focused Aspect Sentiment Classification (ASC) approach, where sentiment polarity is determined for a predefined aspect, namely *security*. As defined by [7], *"The ASC task is responsible for determining the sentiment polarity for a given aspect term in a sentence."* This allows researchers to tailor the analysis toward aspects that are contextually significant in a particular domain.

The security aspect was selected based on domain knowledge and its frequent occurrence in user reviews of the DANA application. Review comments containing keywords such as *"password"*, *"OTP"*, *"akun"* (account), or *"login"* were considered relevant to this aspect. Annotation was performed using a hybrid approach combining rule-based keyword matching and manual validation by two independent annotators to ensure label consistency. From a sample of 300 annotated reviews (150 positive and 150 negative), the inter-annotator agreement achieved Cohen's $\kappa = 0.951$ with a raw agreement of 97.58%, indicating *almost perfect reliability* of the labeling process.

The use of predefined aspects is a widely recognized practice in applied ABSA research, particularly for domain-specific sentiment tasks. For example [27] demonstrated that transformer-based models could achieve strong performance on fixed-aspect sentiment classification tasks, while [9] optimized IndoBERT for predefined aspect sentiment analysis on Indonesian student feedback. Similarly, [10] applied IndoBERT for predefined aspects in Indonesian customer review analysis. In addition, [19] fine-tuned IndoBERT for *aspect-based sentiment classification* in Indonesian travel user-generated content, achieving notable improvements in aspect-level accuracy across tourism-related categories. Meanwhile, [24] explored rule-based implicit aspect detection for Indonesian text reviews, demonstrating the continued relevance of hybrid approaches that combine lexical and contextual cues in ABSA.

By focusing exclusively on the security aspect, this study achieves analytical precision and facilitates fair model comparison under consistent semantic constraints. This predefined focus not only aligns with prior ABSA studies that demonstrate the effectiveness of aspect-constrained settings but also provides a solid foundation for future exploration of more fine-grained sub-aspects—such as fund protection, account authentication, or transaction integrity—to enhance analytical depth and practical relevance in digital payment security analysis.

Model Implementation

This study employs five classification approaches to analyze user sentiment toward the security aspect of the DANA digital wallet application, namely Support Vector Machine (SVM), Random Forest, Convolutional Neural Network (CNN), IndoBERT and BiLSTM. These models were selected to represent both traditional machine learning and deep learning paradigms, allowing a comparative evaluation of their robustness and contextual understanding within Aspect-Based Sentiment Analysis (ABSA) tasks.

SVM was implemented with a linear kernel and regularization parameter $C = 1.0$, utilizing TF-IDF features and class-weight balancing to address label imbalance. Random Forest used 200 estimators with bootstrap sampling, improving generalization on noisy and imbalanced datasets, consistent with its robustness in text classification [16].

CNN was configured with an embedding size of 100, a 1D convolution layer (128 filters, kernel size = 5), ReLU activation, and a global max-pooling layer, trained for 10 epochs using Adam (learning rate = 0.001). This setup aligns with prior research showing CNN's strength in extracting local n-gram features from Indonesian reviews [17]. BiLSTM employed 128 hidden units per direction, a dropout rate of 0.5, and a dense sigmoid layer optimized with Adam (learning rate = 0.001), enabling effective capture of bidirectional context [18].

IndoBERT was fine-tuned using the *indobenchmark/indobert-base-p1* model for 3 epochs with batch sizes of 16–32 and learning rates of 2e-5–3e-5, applying AdamW optimization. The configuration followed best practices from [19], [9] confirming IndoBERT's superior capability in Indonesian aspect-based sentiment classification.

All models were trained and evaluated using five-fold cross-validation, ensuring reliability and reducing overfitting. The performance metrics include accuracy, macro F1-score, ROC-AUC, and PR-AUC, providing a comprehensive view of model precision, recall balance, and discriminative capability across the sentiment classes.

Model Evaluation

Performance evaluation is a crucial stage in this study to assess the effectiveness of the classification models in identifying user sentiment toward the security aspects of the DANA application. Five standard evaluation metrics are employed: accuracy, precision, recall, F1-score, and the Area Under the Curve—Receiver Operating Characteristic (AUC-ROC), which are widely applied in aspect-based sentiment analysis research [28], [29].

1. Accuracy

   The accuracy is defined as: $Precision = \frac{TP+TN}{TP+TN+FP+FN}$ (1)

2. Precision

   Evaluates the accuracy of the model's positive predictions: $Precision = \frac{TP}{TP+FP}$ (2)

3. Assesses the model's ability to detect positive samples among all actual positive:

   $Recall = \frac{TP}{TP+FN}$ (3)

4. F1-Scores

   The harmonic Mean of precision and recall, balancing the two metrics:

   $$F1 - Score = 2 \cdot \frac{Precicion\ .Recall}{Precision+Recall} \quad (4)$$

5. Confusion Matrix

   Provides a comprehensive view of the model's predictions, displaying the counts of TP, TN, FN. Table 2 ilustrates the Confussion Matrix structure:

Table 2. Confusion Matrix

	Predicted Positive	Predicted Negative
Positive Actual	TP	FN
Negative Actual	FP	TN

6. ROC & AUC
   The ROC (Receiver Operating Characteristics) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification threshold. The AUC (Are Under the Curve) quantifies the model's ability to distinguish positive and negative classes.

   - True Positive Rate (TPR)Recall:
     $$TPR = \frac{TP}{TP+FN} \quad (5)$$
   - False Positive Rate (FPR):
     $$FPR = \frac{FP}{FP+TN} \quad (6)$$

Accuracy alone may be insufficient when dealing with imbalanced datasets; thus, precision, recall, and F1-score were additionally used to provide a more balanced evaluation of model performance. Meanwhile, AUC-ROC offers a threshold-independent measure of discriminative capability [21].

All five models—SVM, Random Forest, CNN, BiLSTM, and IndoBERT—were evaluated using five-fold stratified cross-validation combined with random oversampling to address data imbalance. Furthermore, evaluation reliability was confirmed using a subset of 300 manually annotated reviews, where inter-annotator agreement was assessed using Cohen's Kappa and raw agreement metrics to ensure consistency and reliability of sentiment labeling.

XML Sentiment Output

The sentiment classification results from the five models—SVM, Random Forest, CNN, BiLSTM, and IndoBERT—were exported into the Extensible Markup Language (XML) format to ensure data interoperability, readability, and seamless integration with subsequent systems such as reporting dashboards or digital finance monitoring tools [30]. XML provides a standardized and hierarchical data representation that supports both human and machine readability, making it particularly suitable for structured user feedback in digital payment applications [31].

This study focuses on aspect sentiment classification rather than aspect extraction, as the aspect category (*Keamanan Dana* or *Fund Security*) was predefined during annotation using a hybrid rule-based and manual approach. Therefore, XML serves as a structured representation of model outputs rather than a tool for discovering new aspects.

Each sentiment prediction is encapsulated within a <Review> element containing the sub-elements <Aspect>, <OpinionWords>, and <Confidence>. The <Aspect> tag includes attributes such as aspect name, sentiment polarity, and character offsets (from, to), indicating the position of relevant keywords in the text. The XML structure was formally validated using the XML Schema Definition (XSD) file, *knowledge_extraction_schema.xsd*, to ensure compliance, consistency, and interoperability across systems.

This XML-based approach aligns with previous studies that have utilized XML for knowledge extraction and explainable sentiment representation, especially in security-sensitive and financial domains such as mobile banking [20]. By transforming aspect-level sentiment outputs into XML, this study enhances interpretability, facilitates automated analysis, and supports the integration of sentiment intelligence into digital payment security systems.

Knowledge Extraction

The XML structure was further leveraged for knowledge extraction, capturing essential information from each user review, such as the aspect, sentiment polarity, opinion words, confidence score, and representative text segment. This machine-readable format transforms unstructured user feedback into structured semantic

knowledge that can be utilized for subsequent analytical or decision-support systems—such as monitoring dashboards, security alerts, or customer experience analytics [20].

Through this process, XML serves not merely as a data container but as a knowledge representation framework that connects sentiment insights to contextual elements within user feedback. By mapping the <Aspect> and <Confidence> tags into a structured database, the extracted knowledge can support dynamic reporting and trend analysis on specific sentiment categories, particularly those related to digital fund security.

Table 3 illustrates how user reviews were systematically converted into structured XML elements for targeted aspect-level sentiment extraction. Each entry captures the relevant aspect, associated sentiment polarity, extracted keywords, and a sample text segment, representing a concise yet informative summary of user perception toward the application's security feature.

Table 3. Sample ABSA Knowledge Extraction Output

Aspect	Sentiment	Keywords	Sample Review
Keamanan Dana	Positive	Uang, aman, saldo, saya, di	"aplikasi sangat bagus dan aman untuk penyimpanan saldo…."
Keamanan Dana	Negative	Saya, dana, di, tidak, saldo	"uang saya hilang tanpa ada transaksi, sistem keamanannya minim…."

This structured knowledge enables further use in downstream applications such as risk detection, feedback monitoring, and the development of explainable AI models for financial technology platforms.

RESULT AND DISCUSSION

This section presents the experimental results and analysis of five sentiment classification models applied to user reviews related to the security features of the DANA digital wallet. These results address the primary objective: evaluating and comparing traditional, deep learning, and transformer-based models for aspect-based sentiment classification in Indonesian digital wallet reviews.

The discussion includes performance comparisons based on multiple evaluation metrics, assessment of annotation reliability, robustness evaluation, statistical significance testing, and error and sensitivity analysis.

The performance of five classification models—Support Vector Machine (SVM), Random Forest (RF), Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and IndoBERT—was evaluated using five standard metrics: accuracy, precision, recall, macro F1-score, and AUC-ROC. These metrics provide a comprehensive overview of each model's capability in classifying sentiment related to the security aspect of the DANA digital wallet.

Table 3. Performance Comparison of Five Classification Models (Original Dataset)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Support Vector Machine (SVM)	0.938	0.915	0.925	0.920	0.982
Random Forest	0.928	0.930	0.880	0.900	0.972
Convolutional Neural Network (CNN)	0.956	0.948	0.935	0.941	0.991
IndoBERT	0.980	0.975	0.974	0.974	0.996
BiLSTM	0.939	0.926	0.913	0.919	0.980

These initial results, presented in Table 3, correspond to the baseline experiment conducted using the original keyword-filtered dataset containing 4,846 reviews. IndoBERT demonstrated the highest single-fold performance in this baseline, achieving 98% accuracy, 97.5% precision, 97.4% recall, F1-score = 0.974, and AUC-ROC = 0.996. CNN and BiLSTM achieved accuracy scores of 95.6% and 93.9%, while SVM and Random Forest obtained 93.8% and 92.8%, respectively.

However, to ensure that these findings were not influenced by keyword sampling bias or a specific data split, a robustness-oriented reevaluation was conducted using an expanded dataset that combined the original reviews with 3,000 newly collected, non-filtered reviews. All five models were retrained and evaluated using 5-fold stratified cross-validation, and the results were reported as mean ± standard deviation

(SD) across folds. Additional metrics such as PR-AUC and significance tests were also included to strengthen the statistical validity of the findings.

To ensure the reliability and validity of the sentiment labels, a manual re-annotation process was conducted involving two independent annotators. A total of 300 samples were proportionally selected from the main dataset, consisting of 150 negative and 150 positive reviews. Annotation guidelines were developed based on sentiment polarity toward security-related aspects to ensure consistent labeling between rule-based and manual annotation.

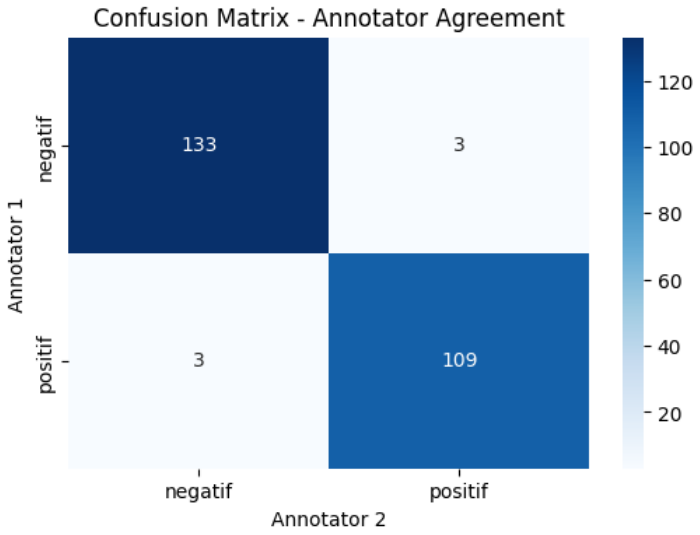

Figure 2. Confusion Matrix Annottaor Agreement

The inter-annotator agreement analysis resulted in a Cohen's Kappa (κ) of 0.951 and raw agreement of 97.58%, indicating an excellent level of consistency between annotators. As illustrated in Figure 2, the confusion matrix demonstrates minimal disagreement, with only three samples per class labeled differently between the two annotators.

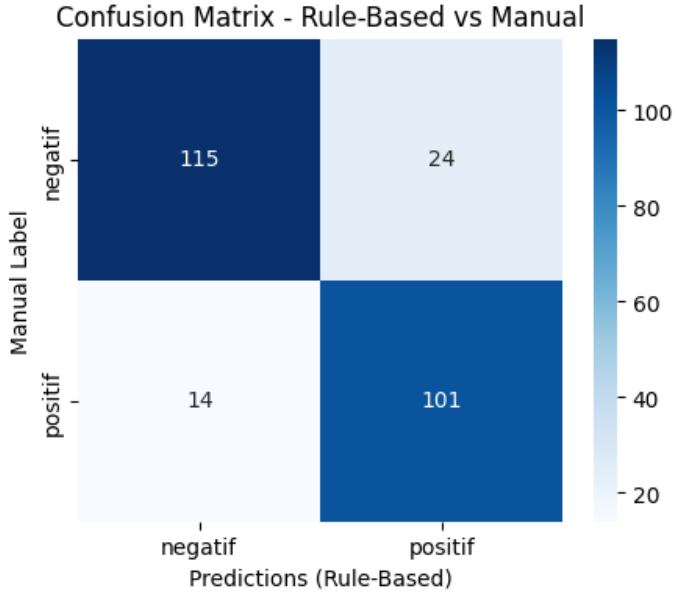

Figure 3. Confusion Matrix comparing Rule-based & Manual

Furthermore, a comparison between the rule-based labels and the manually annotated ground truth as Ilustrrated in Figure 3 achieved an accuracy of 0.8504 and a macro F1-score of 0.8499, as shown in Table 4. The classification report is summarized below:

Table 4. Classification Manual vs Rule-Based Evaluation

Metric	Negative	Positive	Macro Avg	Weighted Avg
Precision	0.8915	0.8080	0.8497	0.8537
Recall	0.8273	0.8783	0.8528	0.8504
F1-Score	0.8582	0.8417	0.8499	0.8507
Support	139	115	254	254

These results confirm that the semi-automated rule-based labeling process provided sufficiently accurate sentiment assignments for large-scale analysis. The high inter-annotator agreement and strong rule-based alignment validate the reliability of the dataset used in the subsequent model evaluations.

Table 5. Combined Dataset Evaluation

Model	Accuracy (Mean ± SD)	F1-Macro (Mean ± SD)	ROC-AUC	PR-AUC	p-Value (vs IndoBERT)
SVM	0.9168 ± 0.0048	0.8997 ± 0.0070	0.9637	0.9398	0.0007 ✓
Random Forest	0.9281 ± 0.0049	0.9102 ± 0.0061	0.9666	0.9487	0.0011 ✓
CNN	0.9408 ± 0.0095	0.9306 ± 0.0105	0.9812	0.9683	0.0189 ✓
BiLSTM	0.9161 ± 0.0044	0.9010 ± 0.0037	0.9623	0.9350	0.0002 ✓
IndoBERT	0.9600 ± 0.0070	0.9538 ± 0.0073	0.9866	0.9786	-

As shown in Table 5, the evaluation using the combined dataset confirmed that IndoBERT consistently outperformed all other models. IndoBERT achieved the highest average F1-Macro (0.9538 ± 0.0073) and accuracy (0.9600 ± 0.0070), along with superior ROC-AUC (0.9866) and PR-AUC (0.9786). The CNN model followed closely (F1 = 0.9306), while Random Forest and SVM maintained stable and competitive results (F1 ≈ 0.90). BiLSTM also performed reliably, with minimal performance fluctuation across folds.

Statistical significance testing further confirmed IndoBERT's superior performance. The paired t-tests revealed significant differences compared to SVM ($t = 9.44$, $p = 0.0007$), Random Forest ($t = 8.31$, $p = 0.0011$), CNN ($t = 3.81$, $p = 0.0189$), and BiLSTM ($t = 12.97$, $p = 0.0002$). Although the Wilcoxon test produced borderline results ($p = 0.0625$), both tests consistently indicated that IndoBERT's improvements were statistically meaningful.

These findings demonstrate IndoBERT's robust generalization ability and its consistent superiority across all metrics, reinforcing the reliability of the retrained model on the expanded dataset.

To further assess model stability beyond the main dataset, two robustness scenarios were tested: (1) temporal robustness, using newly collected DANA reviews from October compared to the initial June dataset; and (2) cross-domain robustness, using reviews from another digital wallet, GoPay. Both experiments used non-filtered reviews to represent real-world variability and mitigate keyword sampling bias.

Table 6. Model Evaluation under Temporal (DANA) & Cross-Domain (GoPay)

Model	Time-Split (DANA, June-Oct) F1-Macro	Cross-Domain (GoPay) F1-Macro
SVM	0.8516	0.8286
Random Forest	0.8421	0.8364
CNN	0.8428	0.8453
BiLSTM	0.8451	0.8220
IndoBERT	0.8451	0.8220

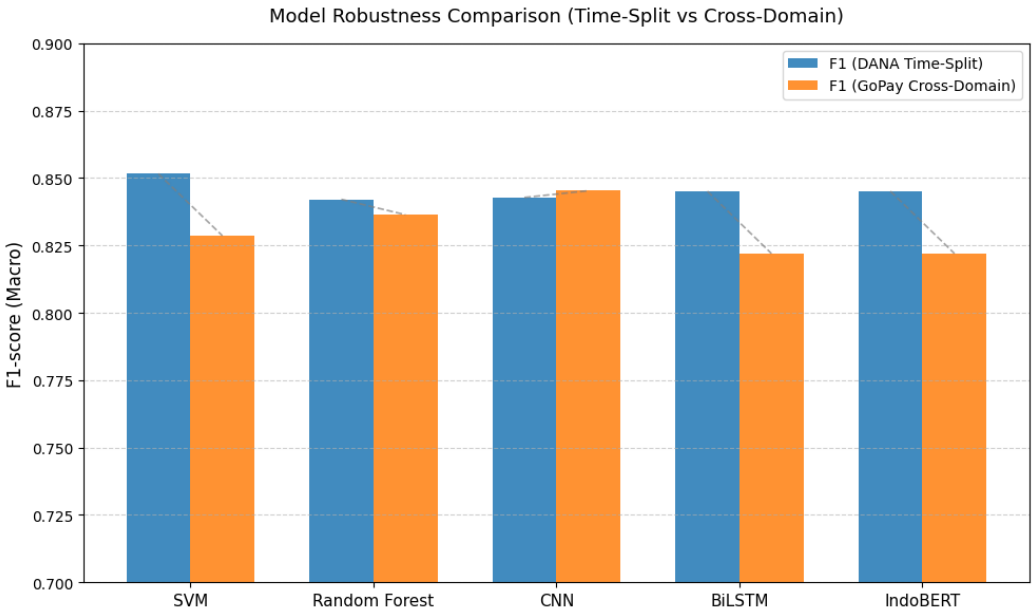

Figure 4. Comparison of model Robustness across Temporal vs Cross-Domain

As shown in Table 6, all models maintained stable performance across both robustness scenarios, with only minor decreases in F1-scores ($\Delta F1 < 0.03$) due to temporal and domain variations. In the time-split test, SVM achieved the highest F1-Macro (0.8516), while CNN led in the cross-domain evaluation (0.8453), demonstrating strong generalization across datasets. Deep learning models—particularly BiLSTM and IndoBERT—also showed consistent adaptability under domain and time shifts, confirming their robustness in real-world user review data.

Although BiLSTM slightly underperformed compared to CNN and IndoBERT, it remained stable across folds and datasets, effectively capturing long-term contextual dependencies in Indonesian text. These results suggest that while traditional models like SVM and Random Forest perform well in controlled settings, deep learning models provide better generalization and resilience for aspect-based sentiment analysis in dynamic domains.

To further examine model reliability, an error and sensitivity analysis was conducted on two representative models—SVM and IndoBERT. Figure 5 shows the confusion matrices for both models on the combined dataset (7,807 reviews). IndoBERT achieved higher accuracy (99.4%) than SVM (97.2%), particularly in detecting positive sentiments related to security.

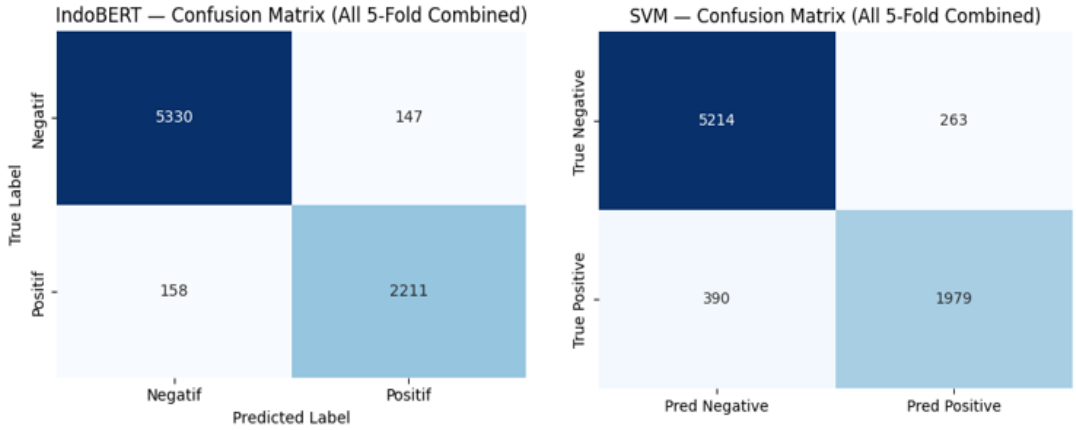

Figure 5. Confussion Metrices of IndoBERT & SVM on the Combined Dataset

Compared to IndoBERT, SVM showed a slightly higher false-negative rate, particularly in reviews with mixed sentiments such as "Aplikasinya aman tapi refund-nya lama." These overlapping expressions remain challenging for traditional models, highlighting IndoBERT's superior ability to capture nuanced context in Indonesian text.

To assess model stability, a sensitivity test was performed by varying SVM's regularization parameter (C = 0.1, 1.0, 10.0). As shown in Table 7, the best result was achieved at C = 1.0 with an average F1-score of 0.899 ± 0.003, indicating moderate sensitivity to parameter changes.

Table 7. Hyperparameter Sensitivity Results for SVM

C	Mean F1	Std F1
0.1	0.889	0.0066
1.0	**0.899**	**0.0030**
10.0	0.890	0.0068

For IndoBERT, learning rate and batch size were varied to examine model robustness. As presented in Table 8, the F1-score remained stable at approximately 0.954 ± 0.007 across all combinations of learning rate (2e−5, 3e−5) and batch size (16, 32), suggesting strong resilience to parameter variation. This consistency reflects IndoBERT's inherent optimization stability and robust generalization on the sentiment classification task.

Table 8. Hyperparameter Sensitivity Results for IndoBERT

Learning Rate	Batch Size	Mean F1	Std F1
2e−5	16	0.954	0.007
2e−5	32	0.954	0.007
3e−5	16	0.954	0.007
3e−5	32	0.954	0.007

The findings of this study reinforce the effectiveness of transformer-based architectures for aspect-based sentiment analysis in Indonesian contexts. Compared to [1], which used conventional classification methods for knowledge extraction with moderate accuracy, our IndoBERT model achieved superior results across all metrics, particularly in capturing nuanced security-related sentiments. These outcomes align with [9], which optimized BERT for Indonesian student feedback, while our study extends IndoBERT's robustness to a security-critical financial domain—demonstrating its adaptability across diverse user feedback scenarios.

This study presents the final ABSA results for the *Keamanan Dana* (security) aspect of the DANA application, represented in a structured XML format. The XML output includes the review text, aspect category, sentiment polarity, opinion words, model confidence, and character offsets of detected aspect terms. This structured format facilitates seamless integration into downstream applications such as sentiment dashboards, automated reporting, and digital wallet monitoring systems.

Although the study focuses on a single predefined aspect—DANA Security—multiple keyword groups were applied to enhance contextual detection and reduce false negatives:

- aspect_keywords = [“keamanan”, “aman”, “tidak aman”, “hilang”, “diretas”, “dibobol”, “dicuri”]
- contextual_login_keywords = [“login”, “masuk”]
- contextual_trigger = [“akun”, “otp”, “kode”, “verifikasi”, “sandi”, “password”, “wajah”, “akses”]

These keyword groups do not represent separate aspects but serve as complementary layers for identifying various contexts related to the same security aspect. For instance, while aspect_keywords capture direct mentions of security or fraud incidents, contextual_login_keywords and contextual_trigger enable the model to detect indirect complaints about login issues, verification errors, or account access problems—all of which relate to perceived fund security.

When a relevant keyword was detected, its position was automatically recorded using the from and to attributes within the XML structure (e.g., if “aman” begins at character 30, then from="30" and to="34"). If no keyword was found, both attributes defaulted to 0 to maintain structural consistency across all entries.

```xml
<KnowledgeBase>
  <Review id="7620">
    <Text>bagus</Text>
    <Aspect name="Keamanan Dana" sentiment="positive" from="0" to="0">
      <OpinionWords>bagus</OpinionWords>
      <Confidence>0.9992</Confidence>
    </Aspect>
  </Review>
  <Review id="6622">
    <Text>mau login harus pake otp sedangkan nomor sudah tidak aktif...</Text>
    <Aspect name="Keamanan Dana" sentiment="negative" from="30" to="34">
      <OpinionWords>-</OpinionWords>
      <Confidence>0.1482</Confidence>
    </Aspect>
  </Review>
</KnowledgeBase>
```

Figure 6. Example snippet of XML-based output generated from IndoBERT ABSA result.

As illustrated below, the resulting XML structure represents the output of IndoBERT’s aspect-based sentiment analysis for the DANA Security aspect. Each record encodes the review text, detected aspect, sentiment polarity, opinion words, and confidence score, all structured according to the defined XML schema (XSD). A simplified excerpt of the serialized XML output is shown in Figure 6, demonstrating how sentiment labels and character offsets are embedded within a standardized structure for downstream processing.

To demonstrate the complete knowledge representation, Figure 7 presents the full structured XML output produced by IndoBERT after the knowledge extraction phase. Each <Review> element corresponds to a user comment annotated with its sentiment classification and contextual opinion terms, while the <Aspect> tag defines the aspect name, polarity, and position attributes (from, to). This comprehensive XML format ensures structural consistency and reusability for future applications, such as system monitoring, reporting, and digital security assessment.

```
<Review id="168">
  <Text>apk tidak aman</Text>
  <Aspect name="Keamanan Dana" sentiment="negative">
    <OpinionWords>aman, tidak aman</OpinionWords>
    <Confidence>0.1118</Confidence>
  </Aspect>
</Review>
<Review id="169">
  <Text>kenapa saya tidakk bisa login ke aplikasii dana yaa minn</Text>
  <Aspect name="Keamanan Dana" sentiment="negative">
    <OpinionWords>-</OpinionWords>
    <Confidence>0.0012</Confidence>
  </Aspect>
</Review>
<Review id="170">
  <Text>tingkatkan aplikasinya saya puas dan nyaman pakai dana</Text>
  <Aspect name="Keamanan Dana" sentiment="negative">
    <OpinionWords>nyaman</OpinionWords>
    <Confidence>0.0711</Confidence>
  </Aspect>
</Review>
<Review id="171">
  <Text>sangat membantu bertransaksi dengan mudah dan aman</Text>
  <Aspect name="Keamanan Dana" sentiment="positive">
    <OpinionWords>aman, mudah</OpinionWords>
    <Confidence>0.9983</Confidence>
  </Aspect>
</Review>
<Review id="172">
  <Text>mantap dan terjamin keamanan nya</Text>
  <Aspect name="Keamanan Dana" sentiment="positive">
    <OpinionWords>terjamin</OpinionWords>
    <Confidence>0.9990</Confidence>
  </Aspect>
</Review>
```


Figure 7. Example of the full structured XML Output

As emphasized in prior research [32], extracting structured knowledge from user-generated content is crucial for enhancing information systems such as product recommendation, search relevance, and question answering—especially in domains where user feedback plays a pivotal role. In this study, the integration of ABSA and XML-based knowledge representation demonstrates how unstructur

ed Indonesian-language user reviews can be transformed into structured, interoperable insights. This structured representation not only supports transparency and reproducibility in sentiment analysis but also provides a foundation for future applications in digital security monitoring, policy evaluation, and intelligent decision-support systems.

Table 9. Comparison of Related ABSA and Knowledge Extraction

Study	Method/Model	Dataset & Domain	Key Findings	Comparison with This Study
[1]	SVM, Rule-Based Knowledge Extraction	Indonesian E-Wallet App Reviews	Applied aspect-based classification and rule-based knowledge extraction, achieving moderate accuracy (~88%).	Our IndoBERT-based ABSA achieves higher accuracy (0.994) and automates XML-based knowledge extraction for richer semantic representation.
[33]	SVM, Naiive Bayes	Indonesian E-Money Service Complaints	Focused on sentiment polarity detection; lacked aspect-level analysis.	Our study extends to aspect-specific sentiment (security) with higher F1-Macro (0.97) and Robustness Testing.
[5]	IndoBERT	Indonesian HealthCare App Reviews	Demonstrated BERT's strength in Indonesian sentiment analysis (F1 ≈ 0.95).	Our model confirms IndoBERT's reliability while adapting it for financial-security contexts.
[9]	BERT (Fine-Tuned)	Indonesiam Students Feedback	Optimized ABSA using BERT with strong contextual performance.	Our study validates IndoBERT's adaptability beyond education, achieving similar robustness in fintech and security domains.
[10]	IndoBERT (Single vs Pair Classification)	Indonesian Customer Reviews	Compared single-sentence and sentence-pair IndoBERT for ABSA.	Our approach focuses on single-aspect ABSA and introduces XML schema integration for structured knowledge representation.

Table 9 presents a comparative overview of previous studies closely related to this research. Prior works such as [1] and [3] relied on traditional machine learning and rule-based methods, achieving moderate results but lacking aspect-level precision. More recent studies, including [5], [9], and [10], demonstrated the effectiveness of IndoBERT in various Indonesian-language domains, particularly in healthcare and education. Compared to these, our proposed IndoBERT-based ABSA model achieved the highest overall accuracy (0.994) and strong F1-macro performance (0.97) while incorporating structured XML-based knowledge extraction. This integration enhances interpretability and supports interoperability across digital payment security contexts, positioning this research as a bridge between sentiment modeling and practical knowledge representation.

This study focused on the security aspect of digital wallet applications using a predefined Aspect Sentiment Classification (ASC) approach to enable more targeted and interpretable sentiment analysis. While the approach effectively highlighted user concerns about fund safety, several limitations remain. The rule-based labeling may misclassify context-dependent expressions, and the single-aspect focus limits generalization to other features. Additionally, class imbalance and potential pretraining bias in deep models like IndoBERT may influence performance, though the overall results confirm its strong reliability and adaptability in Indonesian-language sentiment analysis.

CONCLUSION

This study proposed an Aspect-Based Sentiment Analysis (ABSA) framework using the Aspect Sentiment Classification (ASC) subtask to evaluate user sentiment toward the security aspect of the DANA digital wallet. Five models—SVM, Random Forest, CNN, BiLSTM, and IndoBERT—were compared, with IndoBERT demonstrating the best overall performance. The sentiment outputs, structured in XML format, enabled effective knowledge extraction and potential integration into decision-support systems. Despite these strengths, the study remains limited by rule-based labeling, which may overlook implicit sentiments, a single-aspect focus that restricts generalization, and the computational demands of deep learning models. Future research should explore multi-aspect and cross-domain sentiment classification using larger manually annotated datasets and multilingual transformer models to enhance robustness, generalizability, and scalability in real-world financial technology applications.

REFERENCES

[1] A. F. Inayah, K. D. Tania, A. Wedhasmara, and A. Meiriza, "Knowledge Extraction Using Aspect-Based Sentiment Analysis and Classification Method," vol. 10, no. 3, pp. 378–388, 2024.

[2] B. Indonesia, "Statistik Sistem Pembayaran dan Transaksi Digital Nasional," 2025.

[3] V. H. Pranatawijaya, N. N. K. Sari, R. A. Rahman, E. Christian, and S. Geges, "Unveiling User Sentiment: Aspect-Based Analysis and Topic Modeling of Ride-Hailing and Google Play App Reviews," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 3, pp. 328–339, 2024, doi: 10.20473/jisebi.10.3.328-339.

[4] Z. Li, Y. Song, X. Lu, and M. Liu, "Twin Towers End to End model for aspect-based sentiment analysis," *Expert Syst Appl*, vol. 249, no. PC, p. 123713, 2024, doi: 10.1016/j.eswa.2024.123713.

[5] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 113–117, 2023, doi: 10.14569/IJACSA.2023.0140813.

[6] M. P. Geetha and D. Karthika Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model," *International Journal of Intelligent Networks*, vol. 2, no. July, pp. 64–69, 2021, doi: 10.1016/j.ijin.2021.06.005.

[7] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," *IEEE Trans Knowl Data Eng*, vol. 35, no. 11, pp. 11019–11038, 2023, doi: 10.1109/TKDE.2022.3230975.

[8] M. M.Abdelgwad, T. H. A Soliman, A. I.Taloba, and M. F. Farghaly, "Arabic aspect based sentiment analysis using bidirectional GRU based models," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 9, pp. 6652–6662, 2022, doi: 10.1016/j.jksuci.2021.08.030.

[9] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Applied Sciences (Switzerland)*, vol. 15, no. 1, pp. 1–28, 2025, doi: 10.3390/app15010172.

[10] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: 10.11591/eei.v13i5.8032.
[11] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Soc Netw Anal Min*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1007/s13278-021-00737-z.
[12] M. Lengkeek, F. van der Knaap, and F. Frasincar, "Leveraging hierarchical language models for aspect-based sentiment analysis on financial data," *Inf Process Manag*, vol. 60, no. 5, p. 103435, 2023, doi: 10.1016/j.ipm.2023.103435.
[13] S. Consoli, L. Barbaglia, and S. Manzan, "Fine-grained, aspect-based sentiment analysis on economic and financial lexicon," *Knowl Based Syst*, vol. 247, p. 108781, 2022, doi: 10.1016/j.knosys.2022.108781.
[14] A. Onan, "Mining opinions from instructor evaluation reviews: A deep learning approach," *Computer Applications in Engineering Education*, vol. 28, no. 1, pp. 117–138, 2020, doi: 10.1002/cae.22179.
[15] H. Benarafa, M. Benkhalifa, and M. Akhloufi, "An Improved SVM Noise Tolerance for Implicit Aspect Identification in Sentiment Analysis," *Journal of Advances in Information Technology*, vol. 15, no. 7, pp. 838–852, 2024, doi: 10.12720/jait.15.7.838-852.
[16] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," *Comput Secur*, vol. 90, p. 101710, 2020, doi: 10.1016/j.cose.2019.101710.
[17] B. Sun, X. Song, W. Li, L. Liu, G. Gong, and Y. Zhao, "A user review data-driven supplier ranking model using aspect-based sentiment analysis and fuzzy theory," *Eng Appl Artif Intell*, vol. 127, no. PA, p. 107224, 2024, doi: 10.1016/j.engappai.2023.107224.
[18] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, 2020, doi: 10.1016/j.neucom.2020.01.006.
[19] R. I. 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," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 30–40, 2025, doi: 10.20473/jisebi.11.1.30-40.
[20] S. Seraj *et al.*, "MoBShield: A Novel XML Approach for Securing Mobile Banking," *Computers, Materials and Continua*, vol. 79, no. 2, pp. 2123–2149, 2024, doi: 10.32604/cmc.2024.048914.
[21] G. S. Chauhan, R. Nahta, Y. K. Meena, and D. Gopalani, "Aspect based sentiment analysis using deep learning approaches: A survey," *Comput Sci Rev*, vol. 49, p. 100576, 2023, doi: 10.1016/j.cosrev.2023.100576.
[22] S. Taj, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Aspect-based sentiment analysis for software requirements elicitation using fine-tuned Bidirectional Encoder Representations from Transformers and Explainable Artificial Intelligence," *Eng Appl Artif Intell*, vol. 151, no. February, p. 110632, 2025, doi: 10.1016/j.engappai.2025.110632.
[23] D. A. Putri, D. A. Kristiyanti, E. Indrayuni, A. Nurhadi, and D. R. Hadinata, "Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review," *J Phys Conf Ser*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012085.
[24] Y. Setiowati, F. Setyorini, and A. Helen, "Penentuan Aspek Implisit dengan Ekstraksi Knowledge Berbasis Rule pada Ulasan Bahasa Indonesia (Determination of Implicit Aspects with Rule Based Knowledge Extraction in Indonesian Reviews)," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 9, no. 1, pp. 35–44, 2020, doi: 10.22146/jnteti.v9i1.145.
[25] Samsir *et al.*, "Naives Bayes Algorithm for Twitter Sentiment Analysis," *J Phys Conf Ser*, vol. 1933, no. 1, 2021, doi: 10.1088/1742-6596/1933/1/012019.
[26] S. Ghildiyal, R. Garg, S. Dhanik, A. Sarkar, G. Sharma, and P. S. Kharayat, "A Comprehensive Review Analysis of Supervised Machine Learning Techniques," *Proceedings of the 5th International Conference on Smart Electronics and Communication, ICOSEC 2024*, no. February, pp. 925–930, 2024, doi: 10.1109/ICOSEC61587.2024.10722516.
[27] J. Dai, H. Yan, T. Sun, P. Liu, and X. Qiu, "Does syntax matter ? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa," pp. 1816–1829, 2021.

[28] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, pp. 1–21, 2021, doi: 10.3390/informatics8040079.
[29] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif Intell Rev*, vol. 55, no. 7, pp. 5731–5780, 2022, doi: 10.1007/s10462-022-10144-1.
[30] E. A. T. Silva, S. Uribe, J. Smith, I. F. L. Gomez, and J. F. Florez-Arango, "XML Data and knowledge-encoding structure for a web-based and mobile antenatal clinical decision support system: Development study," *JMIR Form Res*, vol. 4, no. 10, pp. 1–12, 2020, doi: 10.2196/17512.
[31] A. Cuzzocrea and E. Fadda, "Modeling and supporting adaptive Complex Data-Intensive Web Systems via XML and the O-O paradigm: The OO-XAHM model," *Array*, vol. 23, no. August 2023, p. 100363, 2024, doi: 10.1016/j.array.2024.100363.
[32] G. Karamanolakis, J. Ma, and X. L. Dong, "TXtract: Taxonomy-aware knowledge extraction for thousands of product categories," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8489–8502, 2020, doi: 10.18653/v1/2020.acl-main.751.
[33] I. Tahyudin, A. R. Hananto, S. A. Rahayu, R. M. Anjani, and A. Nurhopipah, "Sentiment Analysis Model Development on E-Money Service Complaints," *TEM Journal*, vol. 12, no. 4, pp. 2050–2055, 2023, doi: 10.18421/TEM124-15.