Vol. 7, No. 2, August 2025, pp. 163~175, p-ISSN 2685-2381, e-ISSN 2715-2626
http://dx.doi.org/10.28989/avitec.v7i2.3035

Aspect-Based Sentiment Analysis on User Perceptions of OVO using Latent Dirichlet Allocation and Support Vector Machine

Eka Fahira Aprilia*, Amalia Anjani Arifiyanti, Nambi Sembilu
Department of Information Systems, Universitas Pembangunan Nasional "Veteran" Jawa Timur

Article Info

Article history:
Submitted May 30, 2025
Accepted June 14, 2025
Published June 20, 2025

Keywords:
Aspect-based sentiment analysis;
OVO;
Latent Dirichlet Allocation;
Support Vector Machine.

ABSTRACT

The rapid development of digital technology and the Internet has significantly influenced financial services in Indonesia, leading to the widespread use of digital wallets. One of the most prominent digital wallet platforms is OVO, which has received millions of user reviews across application stores. This study applies aspect-based sentiment analysis to better understand user perceptions from reviews of the OVO application (versions 3.115 to 3.119). A total of 17.086 reviews were collected through web scraping and refined to 4.996 relevant entries. Topic modeling using Latent Dirichlet Allocation (LDA) identified four main aspects frequently discussed by users: Transaction Efficiency, User Experience, Account Access and Registration, and Balance and Charges. However, automatic aspect labeling using LDA keywords achieved only 11.46% agreement with manual annotations, increasing to 40.60% after keyword refinement. Therefore, manual aspect annotation was adopted as the basis for sentiment labeling. Sentiment labeling was conducted by three annotators based on structured guidelines, achieving a Fleiss' Kappa score of 0.9915. A classification model was then developed using the Support Vector Machine (SVM) algorithm across six testing scenarios. The best-performing model, using a Linear kernel without ML-SMOTE, achieved a macro-average precision of 0.843, recall of 0.786, and F1-Score of 0.804. These results demonstrate the model's effectiveness in handling multi-label classification under imbalanced data conditions, particularly for well-distributed aspects such as Transaction Efficiency and User Experience, while highlighting challenges in minority-class detection for aspects such as Account Access and Registration and Balance and Charges.

Corresponding Author:
Eka Fahira Aprilia,
Department of Information Systems, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur, Indonesia
Email: *21082010218@student.upnjatim.ac.id

1. INTRODUCTION

The advancement of technology and the Internet has fundamentally reshaped various dimensions of modern life. One major transformation is evident in the workplace, where many processes have shifted from manual to digital, significantly improving productivity and operational efficiency. Additionally, the Internet has revolutionized human communication through platforms such as social media, email, and instant messaging, making interactions faster and more convenient [1]. It also enables rapid access to diverse sources of information, supporting both learning and decision-making processes [2]. Daily activities such as online shopping, searching for cooking recipes, and accessing health-related content have become easier and more integrated into people's lifestyles. Business practices have also been greatly influenced by the rise of digital platforms, leading to the emergence of e-commerce, digital marketing strategies, and financial technologies (fintech), which continue to shape the global economic landscape [1].

Building upon this transformation, digital technology has also significantly impacted areas such as transportation, communication, shopping, and payment systems. In response to these digital shifts, companies have increasingly adopted technology to meet consumer demands and improve transaction efficiency. As consumers transition from offline to online purchasing behavior, businesses are developing accessible, tech-based payment systems to enhance the convenience and effectiveness of their services [3]. At the national level, Bank Indonesia—the country's central bank—has taken an active role in promoting the shift toward digital financial transactions. One of its key initiatives is the GNNT (National Non-Cash Movement), designed to raise

awareness among the public, businesses, and government institutions about the benefits of using secure, efficient, and practical non-cash payment methods. Despite ongoing developments, the adoption rate of electronic payments in Indonesia remains relatively low compared to other ASEAN countries. Recognizing this potential, Bank Indonesia and the banking sector have collaborated to expand public access to digital financial services through various outreach programs, including online media campaigns [4].

Building upon this foundation, Indonesia has experienced significant growth in digital payment systems, particularly driven by non-bank entities such as transportation platforms and fintech startups. This evolution reflects a transition from the previous dominance of e-money cards issued by commercial banks, signaling a broader transformation in the country's digital financial services landscape [5]. As one of the most prominent branches of financial technology, digital wallets (e-wallets) play a vital role in enabling seamless, user-friendly electronic payment systems in Indonesia [6]. In Indonesia, the popularity of e-wallets continues to grow as they not only support cashless transactions but also assist in personal financial management and broaden access to financial services [7]. As of February 2020, there were 41 e-wallet providers officially licensed by government regulators. The most widely used services include GoPay, OVO, DANA, LinkAja, and ShopeePay. These providers not only facilitate cashless payments but also offer integrated services by collaborating with ride-hailing, food delivery, and entertainment platforms, thereby enhancing the overall user experience and promoting financial inclusion in Indonesia [6].

One of the prominent e-wallet platforms in Indonesia is OVO, which receives a significant volume of user reviews on both the Google Play Store and Apple Store. OVO provides a wide range of digital financial services within a single application, including payment processing, loyalty points, and promotional offers. The application has been installed by over 50 million users, reflecting strong public interest. It currently holds an average rating of 3.8 out of 5 stars on Google Play Store, indicating general user satisfaction, although a notable portion of users have submitted low ratings, highlighting perceived issues and service limitations. These user-submitted reviews offer valuable insights into public perception regarding the application's performance and usability. Common user complaints include frequent transaction failures, security concerns, difficulties during account registration and verification, and unresolved balance refund issues—all pointing to areas requiring improvement. By analyzing these perceptions, developers and service providers can better understand user expectations, address deficiencies, and enhance overall service quality. To support this analysis, aspect-based sentiment analysis can be employed to extract and interpret sentiments associated with specific features or services within the OVO application [8].

To implement this, the research applies ABSA to classify user sentiments based on specific aspects extracted from reviews of the OVO application, from version 3.115.0 to 3.119.0, which were obtained through a web scraping process from the Google Play Store and Apple Store. These versions were selected to ensure the data reflects the most recent user experiences. ABSA itself refers to the task of identifying and classifying sentiments directed at specific aspects within a piece of text. Unlike general sentiment analysis, ABSA enables more granular interpretation by distinguishing different sentiment expressions associated with distinct aspects mentioned in the same sentence or document [9]. Aspects are identified through topic modeling using the Latent Dirichlet Allocation (LDA) method. LDA is one of the most widely used algorithms in topic modeling due to its capability to discover latent topic structures in large, unstructured textual datasets [10].

Previous studies have extensively explored the application of aspect-based sentiment analysis using various classification algorithms such as Support Vector Machine (SVM) to achieve high performance in opinion classification. In study [11], SVM was used to classify sentiment across five hotel-related aspects. The model achieved an average sentiment classification performance of 0.940, demonstrating its effectiveness when combined with semantic similarity techniques and LDA for aspect extraction. Another study applied SVM to reviews of the Flip e-wallet application, focusing on aspects such as transaction speed, security, and fees. The linear kernel configuration yielded the best results, with strong performance in terms of accuracy, precision, and recall across all aspects [12]. Similarly, a comparison of Naive Bayes and SVM for analyzing user sentiment in KAI Access reviews showed that SVM with hyperparameter tuning outperformed other methods, achieving an average accuracy of 91.63%, F1-Score of 75.55%, and precision of 77.60% [13]. These results confirm that SVM is a robust classification method for aspect-based sentiment analysis across diverse domains, supporting its adoption in this research.

Although several prior studies have explored ABSA for e-wallet or transportation applications, many have not addressed the significant challenge of extreme class imbalance in aspect-level sentiment distribution, particularly in multi-label scenarios where one review may cover multiple aspects with different sentiments. This research aims to fill that gap by integrating annotator-based labeling and model evaluation strategies tailored to imbalanced datasets. This study contributes a novel implementation of a multi-label ABSA model using Support Vector Machine (SVM) on OVO application reviews, with performance evaluated across various SVM kernels and oversampling techniques. This study aims to develop and evaluate a multi-label ABSA model for OVO reviews, addressing the challenges of imbalanced aspect distributions.

2. LITERATURE REVIEW

2.1 Aspect-Based Sentiment Analysis

Aspect-Based Sentiment Analysis (ABSA) is an approach in sentiment analysis that focuses on identifying specific aspects of an entity discussed in a text, as well as determining the sentiment associated with each aspect. Compared to general sentiment analysis, ABSA enables a more granular interpretation of opinions by breaking them down into specific components [14]. This method is particularly valuable in domains such as mobile applications and digital wallets, where user reviews often cover various aspects—such as ease of use, security, and service performance—with varying sentiments.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is one of the most widely used algorithms in topic modeling due to its capability to discover underlying topic structures within large collections of unstructured text. As an unsupervised probabilistic model, LDA generates clusters of words that characterize specific topics without relying on predefined labels [10]. Its generative framework is particularly advantageous for handling high-dimensional textual data, where it has demonstrated superior performance compared to alternative topic modeling techniques [15]. Structurally, LDA is a Bayesian model with a three-level hierarchical structure, treating each document as a mixture of latent topics and each topic as a distribution over words [10].

2.3 Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm that is effective for text classification. It was introduced by Vapnik in 1992 as a combination of various concepts in pattern recognition [16]. SVM works by leveraging a hypothesis space of linear functions in a high-dimensional feature space and is trained using optimization algorithms to obtain optimal results [17]. At its core, SVM operates by identifying the most optimal decision boundary, or hyperplane, that separates data into distinct categories. This boundary is defined based on the nearest data points from each class, known as support vectors. By maximizing the margin between the hyperplane and these support vectors, SVM increases the model's confidence in classification and minimizes the risk of misclassification [18]. A major advantage of SVM is its ability to handle overfitting problems, even with high-dimensional feature spaces, and it does not require extensive parameter tuning, as its default parameters are proven to yield strong performance [16].

2.4 Text Pre-processing

Text pre-processing is a fundamental step in natural language processing, aiming to convert unstructured text into a structured and machine-readable format suitable for analysis using machine learning algorithms [19]. This process ensures that the textual data is clean, consistent, and optimized for further computational tasks.

2.4.1 Text Cleaning

This step involves eliminating unnecessary components from the text, such as special characters or irrelevant symbols, and often includes stopword removal and standardizing character formatting to streamline the input for further processing [20].

2.4.2 Case Folding

Words written in uppercase and lowercase are interpreted differently by machines, potentially creating separate vector representations for the same word. To prevent this inconsistency, converting all text to lowercase has become a widely accepted practice in text preprocessing [19].

2.4.3 Tokenization

Tokenization is the process of dividing text into smaller elements known as tokens—such as words, characters, or punctuation—based on spaces or punctuation marks. This segmentation facilitates subsequent filtering words and analysis steps [19].

2.4.4 Normalization

Normalization is applied to convert informal, abbreviated, or misspelled terms into their correct standard forms. This often relies on custom dictionaries to align such terms with a formal vocabulary [21].

2.4.5 Stopword Removal

This technique removes frequently used words—such as prepositions, conjunctions, and pronouns—that typically add little semantic value to the overall analysis [22].

2.4.6 Stemming

Stemming is a technique that trims words to their root forms by removing suffixes. Although it simplifies vocabulary size, it may occasionally alter word meanings and reduce interpretability if applied too aggressively [19].

2.5 Term Frequency-Inverse Document Frequency

Term Frequency–Inverse Document Frequency (TF-IDF) is a statistical technique commonly used in natural language processing, text mining, and information retrieval to quantify the importance of a word in a document relative to a larger corpus [23]. It combines two components: Term Frequency (TF), which measures how often a word appears in a specific document, reflecting its local importance, and Inverse Document Frequency (IDF), which adjusts the weight by reducing the influence of words that commonly appear across many documents and increasing the weight of rarer terms [21]. This approach transforms standardized textual data into meaningful numerical features, allowing machine learning algorithms to interpret and analyze unstructured text effectively [24].

2.6 Confusion Matrix

The confusion matrix is a widely used performance evaluation tool across various scientific and engineering domains, including natural language processing, computer vision, and acoustics. Confusion matrix provides a clear visualization of a classification model's performance by displaying actual versus predicted classifications. For binary classifiers, the matrix is structured into four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), allowing researchers to assess misclassification patterns. This principle can also be extended to multi-class classification, where the matrix highlights common errors between specific classes—offering insights into areas where the model may require refinement or additional distinguishing features [25]. A typical example of a binary confusion matrix is shown in Table 1, which illustrates how predictions are mapped against actual outcomes for positive and negative classes.

Table 1. Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

3. RESEARCH METHODS

This study consists of several stages, starting from problem identification to model evaluation. The research flow is designed to process and analyze data using a text mining and machine learning approach. The systematic research flow is shown in Figure 1.

3.1 Problem Identification

This stage aims to identify issues related to the use of the OVO digital wallet application based on user perceptions of its features, security, and services.

3.2 Literature Study

A literature review was conducted on supporting theories such as topic modeling (LDA) and text classification algorithms (SVM). The literature was obtained from journals, books, and other reliable sources.

3.3 Data Collection

The data used in this study consists of user reviews obtained from the Google Play Store and Apple Store. Web scraping was conducted using the google-play-scraper library for Android reviews and the app_store_scraper library for iOS reviews. The data collection focused on the OVO application version 3.115 until 3.119 and was carried out between August 16, 2024, and October 13, 2024, to capture reviews that reflect the most recent updates and user experiences. All review data were exported and stored in .csv format to facilitate subsequent processing and analysis.

3.4 Text Pre-processing

The collected user review data underwent a series of pre-processing steps to ensure quality and consistency before further analysis. First, duplicate entries and reviews with fewer than two words were removed to eliminate noise. Text cleaning was performed using the re (regular expression) library to strip irrelevant characters such as emojis, numbers, and punctuation marks. Case folding was applied by converting all text to lowercase using Python's built-in text.lower() function. Tokenization was conducted using the word_tokenize function from the nltk library to split the text into individual words. For normalization, an informal-to-formal word dictionary (new_kamusalay.csv) sourced from GitHub was converted into a Python dictionary and applied using a custom normalize() function. Stopword removal utilized the default stoplist from the Sastrawi library, with manual additions based on the frequency analysis of the top 150 most common words. Lastly, stemming was carried out using the Sastrawi library to reduce words to their root forms. These pre-processing steps were performed before topic modeling and classification to ensure a clean and standardized input for analysis.

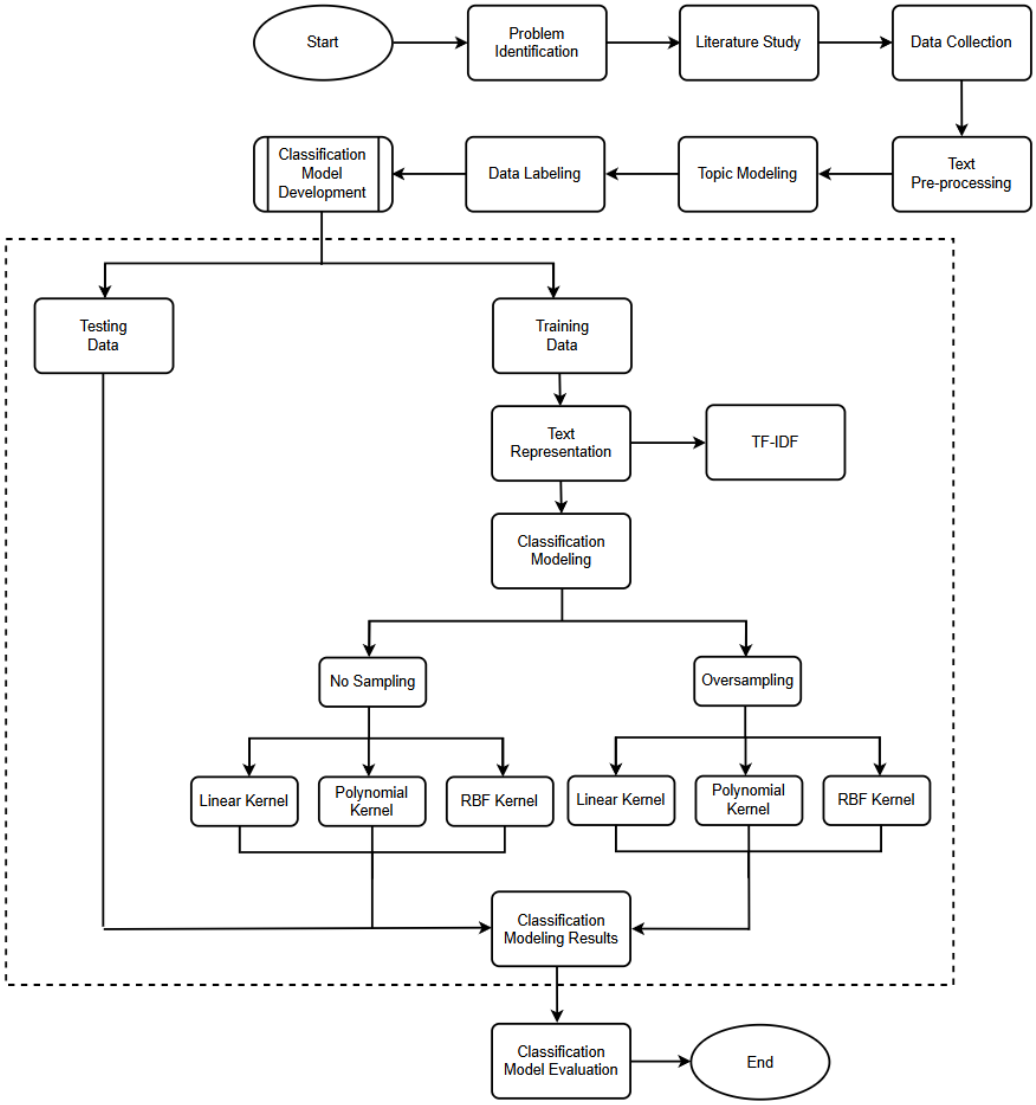

Figure 1. Research systematic flow

3.5 Topic Modeling

Topic modeling in this study was conducted using the Latent Dirichlet Allocation (LDA) method to identify dominant topic present in user reviews. The modeling process utilized the gensim library version 4.3.1 in Python. Before building the model, the textual data was preprocessed and tokenized, then converted into a bag-of-words representation. To determine the most appropriate number of topics, the model was tested across a range of 2 to 10 topics, and the optimal number was selected based on the coherence score results. The model evaluation relied on the coherence score which is commonly used to assess the semantic consistency of the top words within each topic. The model with the highest coherence score was selected as the final topic configuration. The keywords generated from each topic were then manually analyzed and interpreted, enabling the identification of relevant aspects discussed by users in reviews of the OVO application.

3.6 Data Labeling

For aspect labeling, three approaches were applied: automatic labeling using keywords derived from LDA results, automatic labeling with refined LDA keywords, and manual labeling by annotators. Each review was assigned binary labels (1 or 0) based on the presence or absence of aspect-related keywords. After comparing the outcomes of these three approaches, manual labeling was selected as the final reference due to its better contextual accuracy. For sentiment labeling, three annotators were involved and provided with a clear set of guidelines. These guidelines defined four aspects—Transaction Efficiency, User Experience, Account Access and Registration, and Balance and Charges—along with examples of how each aspect could be expressed positively or negatively in a review. Annotators were instructed to assess the sentiment for each identified aspect within a review, assigning positive (1) if the user's expression conveyed satisfaction or success, and negative (2) if it indicated complaints, failures, or dissatisfaction. Each review was labeled independently by all three

annotators. To measure inter-annotator agreement, Fleiss' Kappa was used. Final sentiment labels were determined using a majority voting mechanism to ensure consistency and accuracy in the dataset.

3.7 Classification Model Development

The classification model was developed to predict aspects and sentiments based on user reviews of the OVO application. The data was split using the holdout method with a ratio of 80:20, where 80% was used for training and 20% for testing. The training data was transformed into numerical representations using the TF-IDF method to weight important words in the document. To address class imbalance in multi-label classification, the ML-SMOTE oversampling method was applied in selected scenarios. The classification model was built using the Support Vector Machine (SVM) algorithm with three types of kernels: Linear, Polynomial, and Radial Basis Function (RBF). Each kernel was tested under two conditions: without oversampling and with the application of ML-SMOTE, resulting in six testing scenarios.

To improve model performance, hyperparameter tuning was conducted using Grid Search for all three kernels—Linear, Polynomial, and RBF. The parameter C was tested with values of 0.1, 1, and 10. In addition, the class_weight parameter was set to 'balanced' to address class imbalance. This tuning process aimed to identify the most optimal parameter combination for achieving balanced and accurate classification results across all tested scenarios. The classification model testing scenarios are detailed in Table 2.

Table 2. Classification model testing scenarios

Scenario	Testing Scenario Description	Objective
1	Data split 80:20, SVM with Linear kernel	Evaluate Linear kernel performance on original data without oversampling
2	Data split 80:20, SVM with Linear kernel and ML-SMOTE	Assess ML-SMOTE effect on Linear kernel performance in balanced data
3	Data split 80:20, SVM with Polynomial kernel	Evaluate Polynomial kernel performance on original data without oversampling
4	Data split 80:20, SVM with Polynomial kernel and ML-SMOTE	Assess ML-SMOTE effect on Polynomial kernel performance in balanced data
5	Data split 80:20, SVM with RBF kernel	Evaluate RBF kernel performance on original data without oversampling
6	Data split 80:20, SVM with RBF kernel and ML-SMOTE	Assess ML-SMOTE effect on RBF kernel performance in balanced data

3.8 Classification Model Evaluation

An evaluation was conducted on the six classification model testing scenarios using SVM, which includes three types of kernels (Linear, Polynomial, and RBF), each tested with and without ML-SMOTE. Each model was evaluated using precision, recall, and F1-Score metrics to assess performance on each aspect label. The results of all scenarios were compared to identify the best-performing model. The best model was further analyzed using a confusion matrix to evaluate its effectiveness in classifying the test data comprehensively.

4. RESULTS AND DISCUSSION

4.1 Data Collection

The data in this study was obtained from two main platforms, namely Google Play Store and Apple Store, which provide public reviews from OVO application users. Data collection was carried out using a web scraping technique, extracting the review texts. A total of 7.054 reviews were collected from Google Play Store and 10.032 from Apple Store, resulting in a total of 17.086 reviews. The data was stored in .csv format to facilitate the next stage of analysis.

4.2 Text Pre-processing

Text pre-processing was carried out to clean and simplify user reviews before further analysis. Before entering the main pre-processing stage, initial data cleaning was performed. From a total of 17.086 collected reviews, 9.880 duplicate entries were identified and removed. Then, 1.329 reviews consisting of fewer than two words were eliminated, as they were considered insufficiently informative. After this cleaning process, 5.877 reviews remained.

The remaining reviews were processed through several pre-processing stages. First, text cleaning was performed to remove irrelevant characters such as emojis, numbers, and punctuation. Second, case folding converted all letters to lowercase to standardize text format. Third, tokenizing split sentences into individual words. Fourth, normalization corrected non-standard words to their proper form in accordance with Indonesian language rules. Fifth, stopword removal eliminates common words with low contribution to the meaning of the text. Finally, stemming reduced inflected words to their root form.

After applying all text pre-processing steps, a few reviews were still not properly processed and were excluded from the dataset. The number of reviews after text pre-processing was 5.865. Examples of raw and pre-processed reviews are shown in Table 3.

Table 3. Examples of raw and pre-processed reviews

Raw Review	After Text Pre-processing
great success, very helpful 👍	great success helpful
Sorry, why can't I register my OVO app? I've tried several times and it keeps failing 😳😳😳	sorry app register several fail
Amazing, fast & smooth transaction	amazing fast smooth transaction
Damn, so complicated. From 1 million left only 980 thousand, too much admin fee 😂	damn complicated million left thousand admin fee
This is the third time my top up hasn't been credited since Thursday. Such a trashy e-wallet, causing losses, very bad response. #stopusingovo	third time top up not credited since Thursday e-wallet trash harm people bad response

4.3 Topic Modeling

After text pre-processing, the next stage was topic modeling using the LDA method. The purpose of this step was to identify the main topics in the reviews in order to understand the most frequently discussed aspects. The LDA model was run with the number of topics ranging from 2 to 10, and evaluated using the coherence score metric to determine the most optimal configuration. The evaluation results for each number of topics are presented in Table 4. Based on the evaluation, four topics yielded the highest coherence score of 0.5531 and were selected as the optimal number for further interpretation.

Table 4. Coherence score evaluation for each number of topics

Number of Topics	Coherence Score
2	0.475043
3	0.550549
4	0.553082
5	0.502392
6	0.529926
7	0.508396
8	0.499838
9	0.523013
10	0.528275

The coherence score visualization is shown in Figure 2. The graph indicates that coherence score increases with the number of topics, peaking at four topics. After that, the coherence score declines, indicating that models with more than four topics do not yield better results. Therefore, the LDA model with four topics was used in this study.

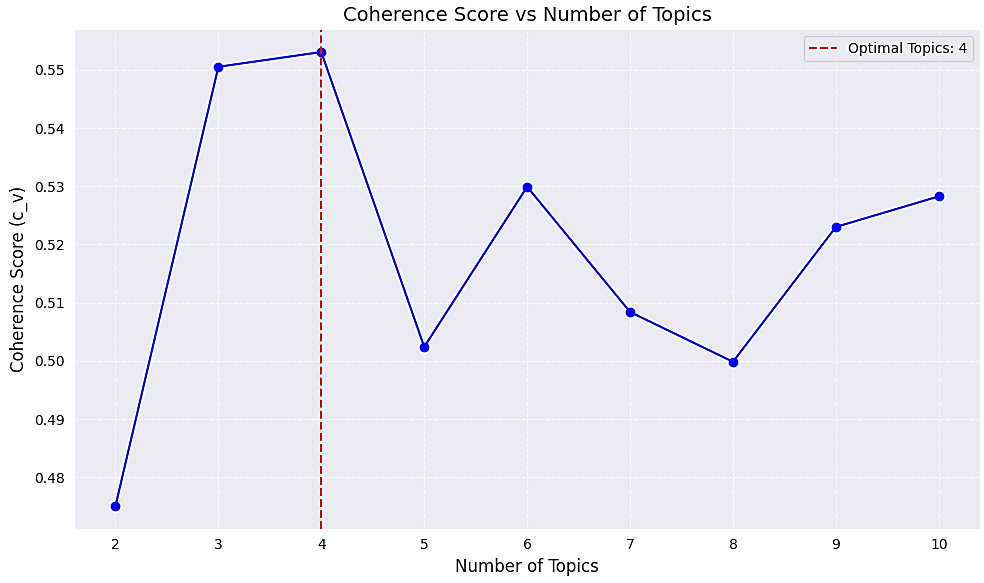

Figure 2. Coherence score visualization for each number of topics

Table 5 presents the top keywords along with their corresponding word probabilities generated for each of the four topics. These probabilities were automatically computed by the LDA model during training and represent the likelihood of each word appearing within a given topic, based on topic-word distributions. The weights were estimated iteratively using a probabilistic generative process, with the implementation provided by the gensim library. By default, gensim applies online variational Bayes inference to optimize the topic-word assignments. As such, the word probabilities were not determined manually or heuristically, but derived directly from the statistical patterns learned from the input corpus. The interpretation of each topic was then manually mapped to a corresponding aspect by analyzing the semantic coherence of its top keywords.

Table 5. Word probabilities for four topics

Topic	Word Probabilities	Aspect Identification
0	0.061*"day" + 0.041*"transfer" + 0.032*"process" + 0.029*"wait" + 0.026*"bank" + 0.021*"great" + 0.019*"money" + 0.016*"more" + 0.015*"work" + 0.014*"entered" + 0.013*"long" + 0.013*"send" + 0.009*"disappointed" + 0.009*"told" + 0.008*"pending"	Transaction Efficiency
1	0.044*"good" + 0.038*"help" + 0.038*"easy" + 0.027*"fast" + 0.026*"transaction" + 0.023*"give" + 0.021*"application" + 0.018*"okay" + 0.016*"thank" + 0.015*"pay" + 0.015*"nice" + 0.014*"more" + 0.013*"make" + 0.012*"use" + 0.011*"wear"	User Experience
2	0.031*"new" + 0.026*"account" + 0.024*"application" + 0.018*"open" + 0.017*"use" + 0.016*"login" + 0.016*"difficult" + 0.016*"enter" + 0.015*"email" + 0.012*"register" + 0.011*"level" + 0.011*"handphone" + 0.010*"complicated" + 0.010*"unclear" + 0.009*"long"	Account Access and Registration
3	0.042*"balance" + 0.029*"transfer" + 0.028*"application" + 0.027*"entered" + 0.025*"deduct" + 0.020*"transaction" + 0.019*"money" + 0.018*"use" + 0.015*"fund" + 0.013*"topup" + 0.012*"pay" + 0.012*"bank" + 0.012*"admin" + 0.011*"process" + 0.010*"fee"	Balance and Charges

4.4 Data Labeling

Data labeling was carried out to identify the aspect and sentiment in each user review, which was then used to build the classification model. Aspect labeling was done using three approaches: automatic labeling based on keywords from LDA results, automatic labeling with improved LDA keywords, and manual labeling by annotators. To determine the most representative approach, the results of automatic labeling were compared against manual annotation as ground truth. The evaluation showed that LDA-based automatic labeling had only 11.46% agreement, while the improved keyword version increased to 40.60%. However, most of the data still lacked context alignment, so manual labeling was selected as it better represented the content of user reviews. Table 6 represents the evaluation results of both automatic approaches compared with manual labeling.

Table 6. Evaluation of automatic vs manual aspect labeling

Labeling Method	Matching Labels	Mismatched Labels	Accuracy (%)
LDA Keyword-Based	573	4.427	11.46%
Improved Keywords	2.030	2.970	40.60%

During the labeling process, reviews that were irrelevant to all four aspects or did not convey clear sentiment were filtered out. These reviews were labeled as “Irrelevant” and excluded from classification model development. Out of the 5.865 reviews that were labeled, 869 were categorized as irrelevant, resulting in 4.996 reviews used for model training and testing.

Sentiment labeling was performed manually by three independent annotators for each labeled aspect. Sentiment was classified into two categories: positive (label 1) and negative (label 2). Inter-annotator agreement was measured using Fleiss’ Kappa, with an overall result of 0.9915, indicating a very high level of agreement. Final labels were determined using a majority voting approach. Examples of labeled reviews for aspect and sentiment are shown in Table 7, illustrating that a single review may contain more than one aspect, each with its own sentiment label.

Table 7. Example reviews with aspect and sentiment labels

Review	Aspect			
	Transaction Efficiency	User Experience	Account Access and Registration	Balance and Charges
The easiest and safest app, makes digital transactions very convenient 🌞 💯 good job OVO app	1	1	0	0
Unclear app, transfer keeps processing and never gets sent, I regret using OVO	2	2	0	0
Tried transferring to bank from 11 PM to 11 AM, still says processing, no response in help center chat	2	2	0	0
My money can't be used because the administration is too complicated	0	0	0	2
Why can't I log into my account, always fails to process	0	0	2	0

Table 8 shows the final distribution of data that has been labeled with aspects and sentiments based on the majority voting results. The results indicate that most reviews carry negative sentiment, particularly in the aspects of Transaction Efficiency, Account Access and Registration, and Balance and Charges. Although the User Experience aspect has a relatively high number of positive sentiments, negative reviews still dominate overall.

Table 8. Final data distribution based on aspect and sentiment

Aspect	Sentiment	
	Positive	Negative
Transaction Efficiency	257	1.507
User Experience	1.382	2.730
Account Access and Registration	8	648
Balance and Charges	27	1.325

4.5 Classification Model Development

The development of the classification model in this study aims to predict aspects and sentiment. The aspects to be classified consist of four categories: Transaction Efficiency, User Experience, Account Access and Registration, and Balance and Charges. Meanwhile, sentiment is classified into two classes: positive (label 1) and negative (label 2). The classification process is carried out in a multi-label setting, where one review can have more than one aspect label, with each aspect assigned its own sentiment label.

To build the classification model, the dataset was split using the holdout method with a ratio of 80:20, where 80% was used as training data and 20% as test data. Feature representation of the text was performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method, which transforms the text into numerical form based on word frequency in the document and the entire corpus. This technique helps the model recognize the most relevant words for each class. The classification model was built using the Support Vector Machine (SVM) algorithm with three different kernel types: Linear, Polynomial, and Radial Basis Function (RBF). Each kernel was tested under two conditions: without and with the application of the ML-SMOTE oversampling technique. ML-SMOTE was applied to address data imbalance for minority labels in specific scenarios. Thus, a total of six model scenarios were evaluated.

In addition, hyperparameter tuning of the parameter C was conducted using GridSearch for each model. The parameter C controls the tolerance for classification errors during model training, making the selection of the optimal value crucial to overall model performance. Each model's performance was evaluated using three main metrics: macro average precision, recall, and F1-Score. The performance of the six models is shown in Table 9.

Table 9. Evaluation results of six test scenarios

Support Vector Machine			Macro Average		
			Precision	Recall	F1-Score
Linear	80:20	Model 1	0.843	0.786	0.804
		Model 2	0.783	0.782	0.783
Polynomial	80:20	Model 3	0.722	0.681	0.696
		Model 4	0.722	0.681	0.696
RBF	80:20	Model 5	0.726	0.718	0.722
		Model 6	0.727	0.719	0.723

4.6 Classification Model Evaluation

The performance evaluation of the classification model was conducted using six testing scenarios involving three types of SVM kernels (Linear, Polynomial, and RBF), each tested with and without the application of the ML-SMOTE oversampling technique. The evaluation metrics used included macro average precision, recall, and F1-Score, as presented in Table 9. These metrics offer a balanced view of the model's performance across all aspect and sentiment labels, especially in a multi-label setting with class imbalance.

Among the six scenarios, Model 1 (SVM with Linear kernel without ML-SMOTE) achieved the best performance, with a macro average F1-Score of 0.804, precision of 0.843, and recall of 0.786. Notably, this strong result was achieved without any oversampling techniques, highlighting the robustness and generalizability of the model. This finding is particularly significant given the multi-label nature of the task and the presence of extreme class imbalance in several aspect categories. It suggests that the TF-IDF representation alone was sufficiently expressive to capture relevant patterns in the review data, rendering additional oversampling unnecessary in this context.

The Linear kernel consistently outperformed both the Polynomial and RBF kernels across all evaluation metrics. The application of ML-SMOTE on the Linear kernel (model 2) slightly decreased performance, indicating that oversampling may have introduced noise rather than improving class balance. The Polynomial kernel (models 3 and 4) produced the weakest results (F1-Score of 0.696), possibly due to its complexity not aligning with the structure of the dataset. The RBF kernel (models 5 and 6) yielded better performance than the Polynomial kernel but remained inferior to the Linear kernel, achieving a highest F1-Score of 0.723. To gain deeper insights into model 1's performance, a confusion matrix was generated for each aspect to examine how accurately the model differentiated between positive and negative sentiment classes, as shown in Figure 3.

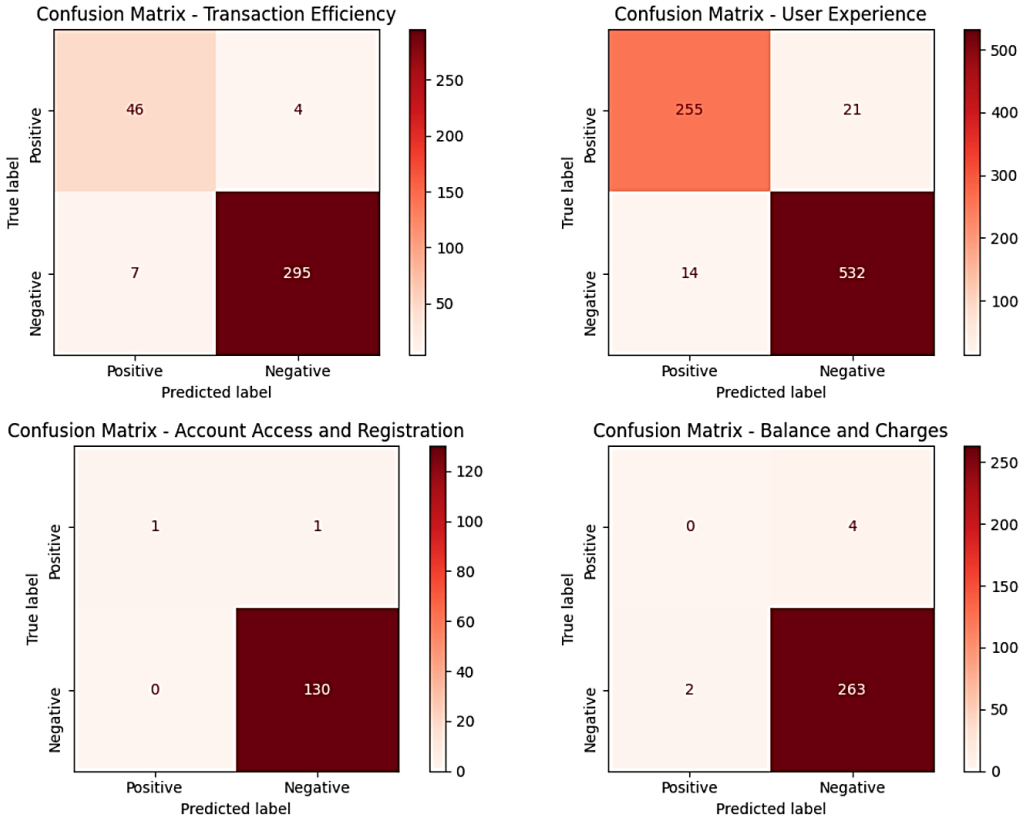

Figure 3. Confusion matrix visualization for each aspect

- Transaction Efficiency

  The model correctly predicted 46 positive reviews and 295 negative reviews. Misclassifications occurred in only 4 positive reviews classified as negative, and 7 negative reviews classified as positive. This demonstrates very good classification performance with minimal errors.
- User Experience

  A balanced and well-predicted aspect, with 255 positive and 532 negative reviews correctly classified. Only 21 positive and 14 negative reviews were misclassified, showing high reliability.
- Account Access and Registration

  This aspect was highly imbalanced. Out of two positive reviews, only one was correctly classified, while the other was misclassified. However, 130 negative reviews were correctly predicted with no false positives.
- Balance and Charges

  Another highly imbalanced aspect. While 263 negative reviews were correctly predicted, none of the four positive reviews were identified. Two negative instances were also misclassified as positive, indicating model bias toward the majority class.

These aspect-specific findings have practical implications for digital wallet providers OVO. For instance, the dominance of negative sentiment in the "Balance and Charges" and "Account Access and Registration" aspects highlights areas that require urgent technical and service improvements. Meanwhile, high user satisfaction in "Transaction Efficiency" and "User Experience" suggests that maintaining performance in these areas can support customer retention and competitive advantage.

Compared to the previous study by [11], which achieved an F1-Score of 0.940 using SVM with semantic similarity and LDA on hotel reviews, the model in this study performed slightly lower. However, it is important to emphasize this study tackles a more complex task involving multi-label classification in a highly imbalanced real-world dataset, without incorporating any semantic enrichment. In contrast to [11], which used single-label classification and predefined aspects, our results demonstrate that a well-optimized Linear SVM can still deliver competitive and practically meaningful performance under more challenging conditions.

In summary, model 1 (SVM Linear without ML-SMOTE) not only achieved the best results among all tested scenarios but also demonstrated the novelty and practicality of employing a straightforward yet powerful approach in a complex setting. This result can be attributed to the high-dimensional and sparse nature of TF-IDF feature vectors, which align well with the strengths of Linear separation in SVM. Unlike Polynomial or RBF kernels that model nonlinear boundaries, the Linear kernel is better suited for text classification tasks with a large number of features and limited training instances per class. These findings highlight the potential of this model for real-world applications such as analysis of e-wallet reviews, enabling service providers like OVO to monitor user sentiment effectively and make informed improvements on aspects. Thus, this study contributes both methodologically—by validating the effectiveness of Linear SVM in complex aspect-based sentiment analysis tasks—and practically by offering actionable insights for service improvement in the fintech industry

5. CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This study successfully developed a multi-label classification model for aspects and sentiment based on user reviews of the OVO application. From a total of 17.086 collected data, filtering and pre-processing were carried out, resulting in 4.996 relevant reviews used to build the classification model. Topic modeling using the LDA method successfully identified four main aspects, namely Transaction Efficiency, User Experience, Account Access and Registration, and Balance and Charges. However, evaluation of automatic aspect labeling based on LDA keyword matching showed a low level of agreement with manual annotation, achieving only 11.46%, and increasing to 40.60% after keyword refinement. These results indicate that LDA is not sufficiently reliable for automatic aspect labeling. Therefore, manual aspect annotation was used as the basis for subsequent sentiment labeling, which was then used to build the classification model. The classification model was built using the SVM algorithm with six testing scenarios. The evaluation results showed that the best model was obtained from the SVM with a Linear kernel without ML-SMOTE oversampling, achieving a precision of 0.843, recall of 0.786, and F1-Score of 0.804. Evaluation using the confusion matrix demonstrated that the model provided good prediction results for aspects with more balanced data distribution, although it still faced challenges in detecting minority classes.

5.2 Recommendation

Future research is advised to balance the data distribution across labels, particularly for minority classes, to achieve more stable and representative classification outcomes. In addition, the text representation approach in this study was limited to the TF-IDF method. For future work, it is recommended to consider more contextual word embedding methods such as Word2Vec, GloVe, or BERT to enhance the semantic quality of information extracted from user reviews. Furthermore, the classification model in this study was built using the SVM algorithm. To obtain more comprehensive results, future studies may compare with other algorithms that are

more adaptive to multi-label scenarios and class imbalance, such as Random Forest, Logistic Regression, XGBoost, or deep learning-based approaches. Although the ML-SMOTE oversampling technique was applied, the results did not significantly improve model performance. Therefore, exploration of other data balancing techniques such as ML-RUS (undersampling), ML-ROS (simple oversampling), or cost-sensitive learning could be potential alternatives to improve the model's sensitivity to minority classes.

REFERENCES

[1] Y. E. Rachmad, A. A. Bakri, R. Nuraini, and T. W. Nurdiani, "Application of the Unified Theory of Acceptance and Use of Technology Method to Analyze Factors Influencing the Use of Digital Wallets in Indonesia," *J. Informasi dan Teknologi*, pp. 229–234, 2024. https://dx.doi.org/10.60083/jidt.v6i1.504

[2] T. W. Nurdiani, "Integrating marketing and finance to increase company performance in VUCA world: a case study on banking state-owned enterprise in Indonesia (MANDIRI, BRI, BTN, BNI)," *European Journal of Business and Innovation Research*, vol. 9, no. 5, pp. 27–32, 2021.

[3] Y. Soelasih, "The factors of millennials' continuance intention to use digital wallets in Indonesia," *Binus Business Review*, vol. 13, no. 3, pp. 315–323, 2022. https://dx.doi.org/10.21512/bbr.v13i3.8561

[4] D. Safitri and M. B. Nainggolan, "Implementation of campaign strategy for national non cash movement from bank of Indonesia," in *Proc. 3rd Int. Conf. on Transformation in Communications (IcoTiC 2017)*, Nov. 2017, pp. 13–17. Atlantis Press. https://dx.doi.org/10.2991/icotic-17.2017.3

[5] R. Widjojo, "The development of digital payment systems in Indonesia: a review of go-pay and ovo e-wallets," *Economic Alternatives*, no. 3, pp. 384–395, 2020. https://dx.doi.org/10.37075/EA.2020.3.03

[6] A. Ciptarianto and Y. Anggoro, "E-Wallet application penetration for financial inclusion in Indonesia," *Int. J. Curr. Sci. Res. Rev.*, vol. 5, no. 2, pp. 319–332, 2022.

[7] I. K. Wati, A. M. Soma, and H. Ispriyahadi, "What Influences User Preferences in Digital Payment Systems? (A Comparative Analysis of E-Wallet in Indonesia)," *Int. J. Entrepreneurship, Business and Creative Economy*, vol. 4, no. 1, p. 78, 2024. https://doi.org/10.31098/ijebce.v4i1.2033

[8] S. D. Widiyaningsih and A. Pertiwi, "Analysis of OVO Application Sentiment Using Lexicon Based Method and K-Nearest Neighbor," *J. Ilm. Ekon. Bisnis*, vol. 25, no. 1, pp. 14–28, 2020.

[9] Y. Noh, S. Park, and S. B. Park, "Aspect-based sentiment analysis using aspect map," *Applied Sciences*, vol. 9, no. 16, p. 3239, 2019. https://dx.doi.org/10.3390/app9163239

[10] Y. T. Lutfi, M. Saputra, and R. Y. Fa'rifah, "Aspect-Based Sentiment Analysis in Identifying Factors Causing Technostress in Fintech Users Using Naïve Bayes Algorithm," in *Proc. Int. Conf. on Enterprise and Industrial Systems* (*ICOEINS 2023*), Dec. 2023, pp. 107–117.

[11] M. D. Pratama, R. Sarno, and R. Abdullah, "Sentiment Analysis User Regarding Hotel Reviews by Aspect Based Using Latent Dirichlet Allocation, Semantic Similarity, and Support Vector Machine Method," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 3, 2022. https://doi.org/10.22266/ijies2022.0630.43

[12] N. Hidayati, F. Hamami, and R. Y. Fa'rifah, "Aspect-Based Sentiment Analysis on FLIP Application Reviews (Play Store) Using Support Vector Machine (SVM) Algorithm," *J. Inform. Telecommun. Eng.*, vol. 7, no. 1, pp. 183–197, 2023.

[13] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indones. J. Comput. Cybern. Syst.)*, vol. 16, no. 2, pp. 113–124, 2022. https://dx.doi.org/10.22146/ijccs.68903

[14] W. A. Awadh, R. B. Sulaiman, and M. A. Mahmoud, "Aspect-based sentiment analysis in MOOCs: a systematic literature review introducing the MASC-MEF framework," *Journal of King Saud University – Computer and Information Sciences*, vol. 37, no. 1, pp. 1–33, 2025. https://dx.doi.org/10.1007/s44443-025-00018-1

[15] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, "Effective comparison of LDA with LSA for topic modelling," in *Proc. 2020 4th Int. Conf. on Intelligent Computing and Control Systems (ICICCS)*, May 2020, pp. 1245–1250. IEEE.

[16] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press, 2007.

[17] D. Irawan, E. B. Perkasa, Y. Yurindra, D. Wahyuningsih, and E. Helmud, "Perbandingan Klasifikasi SMS Berbasis Support Vector Machine, Naive Bayes Classifier, Random Forest dan Bagging Classifier [Comparison of SMS Classification Based on Support Vector Machine, Naive Bayes Classifier, Random Forest, and Bagging Classifier]," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 10, no. 3, pp. 432–437, 2021. (in Indonesian).

[18] F. Ghasemi and S. Sharifi, "Heart Failure Prediction Using Support Vector Machine," *International Journal of Novel Research in Life Sciences*, vol. 12, no. 1, pp. 1–8, 2025.

[19] A. Tabassum and R. R. Patil, "A survey on text pre-processing & feature extraction techniques in natural language processing," *Int. Res. J. Eng. Technol. (IRJET)*, vol. 7, no. 6, pp. 4864–4867, 2020.

[20] B. Pahwa, S. Taruna, and N. Kasliwal, "Sentiment analysis-strategy for text pre-processing," *Int. J. Comput. Appl.*, vol. 180, no. 34, pp. 15–18, 2018. https://doi.org/10.5120/ijca2018916865

[21] A. R. Royyan and E. B. Setiawan, "Feature expansion Word2Vec for sentiment analysis of public policy in Twitter," *J. RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, pp. 78–84, 2022. https://doi.org/10.29207/resti.v6i1.3525

[22] S. E. Saad and J. Yang, "Twitter sentiment analysis based on ordinal regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019. https://dx.doi.org/10.1109/ACCESS.2019.2958804

[23] Z. Yang, X. Wang, M. Qiu, S. Hou, and Y. Wu, "Account of Spatio-Temporal Characteristics in Customs Anti-Smuggling Intelligence Acquisition: A Combined LSTM+ CRF Model Using TF-IDF and Levenshtein," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 20, no. 1, pp. 1–20, 2024.

[24] E. Chan, "The Most Optimal Machine Learning Model for Defense Stock Prediction," *Curieux Academic Journal*, vol. 3, no. 2, pp. 45–56, 2024.

[25] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," *Computer Science & Information Technology*, vol. 1, pp. 1–14, 2020.