Vol. 7, No. 2, August 2025, pp. 209~222, p-ISSN 2685-2381, e-ISSN 2715-2626
http://dx.doi.org/10.28989/avitec.v7i2.3033

A Multilingual Approach to Aspect-Based Sentiment Analysis on Gobis Suroboyo Application Reviews using LDA and SVM

Dianita Puspitasari*, Eka Dyar Wahyuni, Reisa Permatasari
Department of Information Systems, UPN Veteran Jawa Timur, Indonesia

Article Info

Article history:
Submitted May 29, 2025
Accepted June 13, 2025
Published June 24, 2025

Keywords:
Aspect-Based Sentiment Analysis;
Latent Dirichlet Allocation;
Support Vector Machine;
multilingual;
GOBIS.

ABSTRACT

The GOBIS application, developed by the Surabaya City Transportation Department, is a digital service designed to provide public transportation information and reduce traffic congestion. Despite having exceeded 100,000 downloads, the application has received numerous complaints from users, as reflected in the multilingual reviews on its platform. To ensure analytical consistency, this research focuses solely on reviews in Indonesian and English. Using Aspect-Based Sentiment Analysis (ABSA), this study employs Latent Dirichlet Allocation (LDA) for aspect identification and Support Vector Machine (SVM) for sentiment classification. The aim of this research is to determine the dominant aspects in user feedback and evaluate the effectiveness of the Support Vector Machine (SVM) model in classifying multilingual reviews. The research results show six main aspects that frequently appear in reviews, namely Application Features and Development, User Suggestions and Service Innovation, Error and Location Accuracy, Delay and Application Usability, Comfort and Service Quality, as well as Route Tracking and Vehicle Information. The Support Vector Machine (SVM) model, tested with 10-fold cross-validation, demonstrates consistent performance, achieving balanced metrics accuracy (74.16%), precision (73.76%), recall (73.54%), and F1-score (73.63%). This highlights its capability in handling multilingual sentiment analysis for application improvement.

Corresponding Author:
Dianita Puspitasari,
Department of Information Systems, UPN Veteran Jawa Timur,
Jl. Raya Rungkut Madya, Gunung Anyar, Surabaya, Indonesia.
Email: *dianitapuspitasari27@gmail.com

1. INTRODUCTION

Surabaya, the capital of East Java Province, is among the most densely populated cities in Indonesia. According to data from the Central Statistics Agency (*Badan Pusat Statistik*, abbreviated as BPS) in 2023, the city's population reached 3,009,286 people [1]. This rapid population growth has brought about several challenges, particularly in the transportation sector. One of the most pressing issues is the community's strong reliance on private vehicles, such as motorcycles and cars. This heavy dependence has led to an imbalance between the volume of vehicles and the capacity of the existing road infrastructure, resulting in worsening traffic congestion, where vehicle buildup frequently causes significant delays and road blockages [2].

In an effort to alleviate traffic congestion and reduce air pollution, the Surabaya City Government introduced a more environmentally friendly mode of public transport known as the Suroboyo Bus. Officially launched on April 7, 2018, by the Surabaya City Transportation Department, the bus is 12 meters long and 2.4 meters wide [3]. To support this initiative, the Regional Technical Implementation Unit (*Unit Pelaksana Teknis Daerah*, abbreviated as UPTD) for Public Transport Management developed a digital application called GOBIS (*Golek Bis*, meaning "find a bus" in Javanese), designed to provide comprehensive information about Surabaya's public transport services. The GOBIS Suroboyo application offers access to services such as the Suroboyo Bus, Feeder Wirawiri, Teman Bus, and Trans Jatim Bus. It also includes several features, such as balance top-up, an integrated real-time route map, a bottle exchange post for plastic recycling, and a Frequently Asked Questions (FAQ) section.

With over 100,000 downloads by Android users across Indonesia, the application has received a wide range of user feedback. These reviews highlight not only challenges experienced by users but also appreciation for the application's features and services. Notably, the reviews are written in various languages, including Indonesian, Javanese, English, and mixed-language formats. This linguistic diversity potentially offers deeper

insight into user satisfaction. However, it also presents challenges in data management and analysis, particularly as the volume of feedback continues to increase. Since manually reviewing the data is inefficient, this study focuses solely on reviews written in Indonesian and English. Indonesian was selected as the primary language used by most GOBIS users, while English was included due to its common use in application reviews internationally. This limitation is intended to improve the relevance of the data and support a more effective sentiment analysis process.

This research adopts the Aspect-Based Sentiment Analysis (ABSA) method, which involves two primary stages aspect identification and sentiment classification [4]. In the aspect extraction phase, a Topic Modeling technique is applied using the Latent Dirichlet Allocation (LDA) algorithm, which probabilistically uncovers hidden thematic structures within user reviews [5]. For sentiment classification, the Support Vector Machine (SVM) algorithm is employed to categorize reviews and evaluate the model's effectiveness in identifying sentiment. The sentiment polarity, whether positive or negative, is automatically determined using a lexicon-based approach integrated with SVM, which relies on a sentiment dictionary for labeling [6].

Several previous studies have applied Aspect-Based Sentiment Analysis (ABSA) in the service sector, particularly in transportation and tourism. Airin et al. [7] implemented ABSA on e-hailing services in Malaysia using Latent Dirichlet Allocation (LDA) for aspect extraction and Support Vector Machine (SVM) for sentiment classification. Their study evaluated topic quality using both perplexity and coherence scores and achieved relatively strong classification performance in English (F1-score 88.0%) and Malay (F1-score 70.0%). However, the multilingual datasets were processed separately per language, with distinct preprocessing pipelines, and without explicit language identification or translation. This limits the applicability of their approach to more realistic code-mixed user reviews, which are common in mobile application contexts. Moreover, their datasets were collected from general repositories and social media, not from a single domain-specific application.

Meanwhile, Mustakim and Priyanta [8] analyzed reviews from the KAI Access application using supervised learning with manually labeled aspects. While their study utilized SVM, it was limited to monolingual Indonesian reviews and did not apply unsupervised techniques such as LDA for discovering latent aspect structures. On the other hand, Andono et al. [9] combined LDA with BERT and semantic similarity to analyze hotel reviews, achieving high classification performance (F1-score 0.97). However, their research was confined to the hospitality domain, did not explicitly address multilingual or code-mixed inputs, and lacked a fully integrated ABSA pipeline encompassing preprocessing, aspect modeling, and sentiment classification.

In contrast to these studies, the present research focuses on real-world reviews from a local, community-based public transportation application, namely GOBIS Suroboyo, which has not been previously studied. This study introduces a comprehensive multilingual preprocessing pipeline, including language identification, slang normalization, synonym substitution, and machine translation to unify Indonesian-English code-mixed reviews into a consistent representation. Aspect extraction is performed using LDA, with topic quality evaluated through both coherence and stability scores, while sentiment classification is implemented using SVM, with all stages embedded in a fully integrated and reproducible ABSA pipeline.

Through this approach, this study aims to provide a novel contribution to Aspect-Based Sentiment Analysis (ABSA), particularly in the context of local public transportation and the processing of multilingual user reviews, as well as to generate data-driven recommendations to improve the quality of the GOBIS Suroboyo application services.

2. RESEARCH METHODS

This section delineates the systematic approach employed in this study to ensure a structured investigation and the attainment of the anticipated outcomes. Each phase of the methodology, from the data collection to model validation, is meticulously designed to uphold the integrity, accuracy, and validity of the data collection and analysis processes. Figure 1 illustrates the sequential stages of this study's methodology, encompassing data collection, data exploratory, preprocessing, topic modeling, labeling, data splitting, model development, and model validation.

2.1 Data Collection

The data collection process for this study involved extracting user reviews of the GOBIS Suroboyo application from the Google Play Store. To achieve this, the google-play-scraper Python library was utilized, which facilitates the automated retrieval of app-related data, including user reviews, ratings, and other metadata. The scraping process targeted reviews posted between April 7, 2018 the official launch date of the GOBIS Suroboyo application and October 2024. This timeframe was selected to encompass the full range of user feedback available up to the most recent data prior to the study's commencement. Upon successful extraction, the collected data were stored in Comma-Separated Values (CSV) format. This format was chosen for its compatibility with various data analysis tools and its ease of use in subsequent preprocessing and analysis stages.

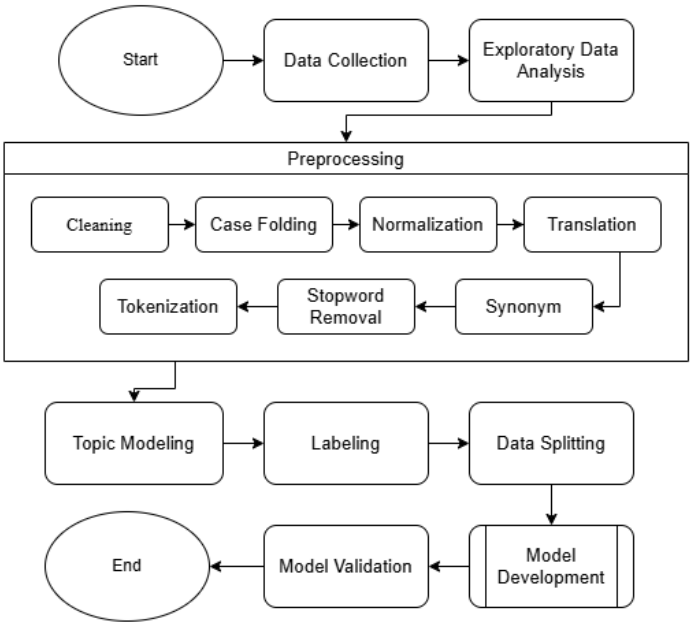

Figure 1. Methodology flowchart

2.2 Exploratory Data Analysis

In this phase, the study conducted an analysis of the average word count per review to understand the typical length of user feedback and identify any excessively long reviews that might affect the consistency of sentiment analysis results. Language detection techniques were applied to each review to classify and group the data based on the language used, focusing primarily on Indonesian and English. To visualize the most frequently occurring terms and gain insights into common themes within the reviews, wordclouds were generated separately for the Indonesian and English subsets of the data.

2.3 Preprocessing

To prepare the multilingual user reviews for effective sentiment analysis, a comprehensive preprocessing pipeline was implemented, encompassing several critical steps to clean and standardize the textual data:

1. Cleaning is a preprocessing step that usually involves removing punctuation, numbers, non-ASCII special characters, URLs, and excessive white space [10]. Additionally, to maintain language consistency, reviews written in languages other than Indonesian or English are deleted. Duplicate entries are identified and removed to prevent redundancy.
2. Case folding is a text preprocessing technique that involves converting all characters to lowercase to ensure uniformity and reduce feature dimensionality [10].
3. Normalization refers to the process of transforming informal terms or abbreviated expressions into their standardized forms based on the guidelines of the Big Indonesian Dictionary (*Kamus Besar Bahasa Indonesia*, abbreviated as KBBI) [11]. For English-language reviews, terms were retained in their original form at this stage.
4. Translation: To streamline subsequent analysis, all reviews were translated into English, facilitating a unified processing approach and leveraging English-based analytical tools [12].
5. Synonym: Synonyms are the process of eliminating features (words) that have similar meanings, aimed at making the data dictionary denser (important features) so that the distribution in the word vector is not too broad [13].
6. Stopword removal involves filtering out high-frequency but low-meaning words such as “and”, “which”, and “in” since they typically contribute little value to text classification tasks [14]. However, certain words deemed important for sentiment analysis, such as “no”, “error”, and “broken”, were retained to preserve critical contextual information
7. Tokenization: The cleaned and standardized text was then tokenized, breaking down the sentences into individual words or tokens [4] [15]. This step is essential for converting the text into a format suitable for machine learning algorithms and further analysis.

This preprocessing framework ensured that the textual data was clean, consistent, and appropriately formatted, laying a solid foundation for the subsequent sentiment analysis tasks.

2.4 Topic Modeling

Latent Dirichlet Allocation (LDA) is a probabilistic topic modeling technique widely used in text analysis to uncover latent themes within large textual corpora [16] [17]. In this study, LDA was employed to identify dominant topics from user reviews of the GOBIS Suroboyo application. The model was implemented using the Gensim library, with key hyperparameters defined to ensure reproducibility.

To determine the optimal number of topics, a combined evaluation of coherence scores (c_v) which consider word co-occurrence and semantic similarity and stability scores was conducted [18]. Several topic counts were tested, and the configuration that yielded the most coherent and interpretable results was selected.

Once the optimal number of topics was established, the LDA model was applied to the preprocessed corpus. The algorithm identifies word co-occurrence patterns across documents, infers topic distributions per document, and assigns probability scores to each word within a topic. This process results in clusters of semantically related words, which were manually interpreted and labeled based on their most representative keywords. The resulting labeled topics provide meaningful insights into user perceptions and highlight key areas of concern or satisfaction within the application.

The use of LDA in this study was chosen due to its effectiveness in handling large-scale, unlabeled textual data, which is common in public digital reviews. Unlike rule-based approaches which require predefined aspect dictionaries or linguistic rules LDA automatically uncovers latent themes without the need for explicit aspect preparation [19][20]. Furthermore, research by Ozyurt & Akçayol [21], which introduced the SS-LDA (Sentence-Segmented LDA) model, demonstrated that LDA applied to sentence-level segments can effectively identify aspect groupings, even in short texts, without manual annotation. This method is also highly adaptable across different domains without requiring refinement of aspect lexicons.

In addition, compared to supervised approaches such as multi-label classification or neural networks, LDA offers efficiency advantages by eliminating the need for complex and costly data labeling processes. A study by Ozyurt & Akçayol [21] showed that in low-resource settings, LDA was able to generate semantically coherent and diverse aspect groupings, even outperforming certain supervised models trained on limited data. Therefore, in the context of early-stage exploration or when labeled datasets are unavailable, LDA remains a relevant and effective approach for aspect-based sentiment analysis.

2.5 Labeling

In this phase, aspect and sentiment labeling is carried out to facilitate aspect-based sentiment analysis. Aspect labeling uses topics identified through Latent Dirichlet Allocation (LDA) as a reference, which assists in classifying user reviews into relevant aspects. This process involves linking each review with the dominant aspect based on keywords commonly identified by LDA. Sentiment labeling aims to categorize each review as positive or negative. This is achieved using a lexicon-based approach, referring to a sentiment dictionary to assess the polarity of words in the reviews [22]. The lexicons used are TextBlob and Pattern, as both are lightweight, easy-to-use lexicon-based libraries that can produce sentiment polarity scores in the form of continuous values (ranging from -1 to 1) reflecting the intensity of sentiment in a text. Additionally, Pattern also provides information on the level of subjectivity, which can enrich the analysis results. The selection of these two lexicons is also based on previous studies that demonstrated stable and efficient performance across various types of data [23]. Additionally, manual labeling is also performed for sentiment labeling. After the sentiment labeling process is completed, the next step is to measure the level of agreement among raters using Fleiss' Kappa. Fleiss' Kappa is a method used to evaluate the consistency or agreement among several annotators in labeling data, especially when involving more than two raters[24]. This measurement is important to ensure that the labeling is reliable and consistent, especially in the context of assessments carried out independently by multiple annotators.

2.6 Data Splitting

In this phase, the dataset was divided using the holdout method with a 90:10 ratio, allocating 90% of the data for model training and 10% for validation. This approach ensures that the model is trained on a substantial portion of the data, enhancing its ability to learn underlying patterns, while retaining a separate validation set to evaluate its performance on unseen data. The split was performed randomly to minimize bias and ensure that both subsets are representative of the overall dataset.

2.7 Model Development

The model development phase encompasses several critical steps to build and evaluate a sentiment classification model for user reviews of the GOBIS Suroboyo application. The diagram in Figure 2 represents the flowchart of model development. This process includes data separation for modeling purposes, term weighting using the TF-IDF method, classification process with the Support Vector Machine (SVM) algorithm, and evaluation of the model's performance.

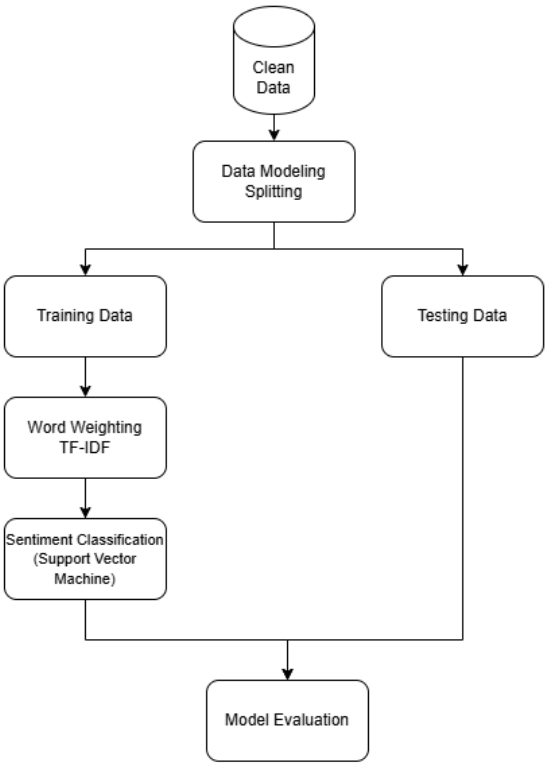

Figure 2. Model development flowchart

2.7.1 Data Splitting for Modeling

In this study, multiple data splitting strategies were employed to evaluate the performance and generalizability of the sentiment classification model. Specifically, the holdout method was applied with three different training-to-testing ratios: 90:10, 80:20, and 70:30. Additionally, k-fold cross-validation was utilized with k values of 5 and 10, allowing each subset of the data to serve as both training and testing data across different iterations. These approaches help in assessing the model's performance and generalizability.

2.7.2 Term Weighting with TF-IDF

The Term Frequency-Inverse Document Frequency (TF-IDF) method is utilized to assign weights to words based on their frequency within a document and their rarity across the entire corpus [6]. This approach emphasizes terms that are significant to a specific document while diminishing the influence of commonly occurring words.

The calculation of TF-IDF involves the following steps:

1. Term Frequency (TF): Quantifies how often a term appears in a document.
2. Inverse Document Frequency (IDF): Assesses the rarity of the term across all documents in the corpus.
3. TF-IDF Score: Combines TF and IDF to assign a weight that reflects the term's importance within the document.

2.7.3 Sentiment Classification Using Support Vector Machine (SVM)

In this study, the sentiment classification phase employs Support Vector Machine (SVM) to categorize user reviews of the Gobis Suroboyo application into positive or negative sentiments. Following the application of Term Frequency-Inverse Document Frequency (TF-IDF) for word weighting, the processed review data serve as input for the SVM model. Support Vector Machine (SVM) is a supervised learning algorithm used for classification tasks, where the model learns to distinguish between categories by being trained on labeled datasets, which are typically divided into training and testing sets [25] [26]. SVM works by identifying an optimal decision boundary referred to as a hyperplane that separates data points of different classes as distinctly as possible. The core idea behind SVM is to maximize the margin between the hyperplane and the closest data points from each class, known as support vectors, which contributes to improved generalization performance. During the training phase, the algorithm iteratively searches for the hyperplane that achieves the widest possible margin while minimizing classification errors on both the training and testing data [25]. For this study, sentiment classification is conducted using the LinearSVC implementation from the scikit-learn library, chosen for its efficiency and suitability in handling high-dimensional sparse text data produced by the TF-IDF representation.

The selection of SVM in this study is also supported by previous research that has demonstrated its superior performance compared to other classification algorithms such as Naive Bayes and k-Nearest Neighbor.

For example, Mustakim and Priyanta [8], in their study on aspect-based sentiment analysis of KAI Access application reviews, compared the performance of Support Vector Machine (SVM) and Naive Bayes Classifier (NBC). Their results showed that SVM with hyperparameter tuning achieved the highest accuracy of 91.63% and an F1-score of 75.55%, significantly outperforming NBC, which only reached 83.47% accuracy and an F1-score of 61.92%. Similarly, a study by Iskandar and Nataliani [19] that compared three classification algorithms SVM, Naive Bayes, and k-NN for aspect-based sentiment analysis on gadget reviews reported that SVM achieved the highest average accuracy of 96.43%, outperforming both Naive Bayes (83.54%) and k-NN (59.68%). Based on these findings, SVM is considered more reliable and consistent in handling sentiment classification tasks involving complex and imbalanced textual data, making it a suitable choice for this research.

2.7.4 Model Evaluation

Upon completing the classification process, the subsequent phase involves assessing the model's performance. This evaluation determines how effectively the model categorizes aspects and sentiments within each review. To quantify the model's efficacy, standard evaluation metrics such as accuracy, precision, recall, and F1-score are employed. These metrics provide insights into the model's ability to make correct predictions and handle various data scenarios.

2.8 Model Validation

Following the completion of the model evaluation, the next step is to validate the model using a separate dataset that was set aside during the initial data preparation phase. This validation ensures that the model performs effectively in real-world scenarios and maintains its accuracy when applied to new data. The validation data undergo the same preprocessing steps as the training data, including text normalization and TF-IDF weighting, to ensure consistency. Once prepared, the validation dataset is input into the trained model, and its performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to classify sentiments correctly and generalize to unseen data.

3. RESULTS AND DISCUSSION

In this study, the Python programming language was used, utilizing Google Colab as the analysis platform. Based on the literature review, the Support Vector Machine (SVM) method has proven effective in classifying review texts with a high accuracy rate. Furthermore, the application of the Latent Dirichlet Allocation (LDA) model optimizes the identification of aspects within the text data. Meanwhile, the Lexicon-Based Labeling technique speeds up the sentiment analysis process by utilizing available lexicon dictionaries or libraries.

3.1 Data Collection

The initial stage in this research process begins with data collection, focusing on gathering user reviews of the GOBIS Suroboyo application. Data was collected by extracting reviews from the Google Play Store using the google_play_scraper library in Python. The Google Play Store was chosen as a source because it provides direct feedback from users of the GOBIS Suroboyo application, which is a crucial basis for further analysis.

reviewId	userName	userImage	content	score
f353e5c9-59b1-4	22-Yusuf Qardhc	https://play-lh.go	Orang cuma mau liat posisi bus malah di	5
5d8fc325-8eae-4	Muhammad Arfa	https://play-lh.go	Aplikasi updatenya terbaru sudah seperti	5
7e09ef3a-6f29-4	hn1231	https://play-lh.go	Semoga jumlah armada makin banyak se	4
01539bfb-9d52-4	steven beta	https://play-lh.go	Sudah login msh di suruh login lagi	1
08bbac1d-87ec-	Rizkyary Prasety	https://play-lh.go	Apaan sih, udah diundang malah gabisa	1
97f026e2-7aa9-4	Daniel Adi	https://play-lh.go	Apk ini mengumpulkan data2 pribadi sen	1
f2dba353-93d0-4	Taqiyuddin Ja'far	https://play-lh.go	Dipaksa update, harus login. Sudah login	1
c4ceed83-186e-	Dina Ms	https://play-lh.go	Sebelum di update untuk track feeder ma	1
f839d0fa-14c5-4	hanifa novitasari	https://play-lh.go	Untuk segi pelayanan dan transportasi ny	3
4c7a2a23-3ea8-	m ganda abdi wi	https://play-lh.go	Aplikasinya kurang sekali, perlu ditingkat	1
0770e968-457b-	Siti Hatijah	https://play-lh.go	sangat senang	5
4c7c5053-9169-	The Sun	https://play-lh.go	Sebuah kemunduran dari aplikasi gobis,	1
1dd6a3fd-be4b-4	Ann Hermione	https://play-lh.go	Aplikasinya lemot bgt, gak bisa muncul b	1
ae96f5b4-a3e0-4	Zulfa Aulia	https://play-lh.go	ga guna, buat cek rute bis tapi bis nya aja	1

Figure 3. Result of data scraping

The Figure 3 shows the results of the data scraping process, which successfully collected 1,168 reviews along with 11 columns of information. However, in this study, only the content column is used for analysis, because that column contains the review text which is the main focus of the aspect-based sentiment analysis.

3.2 Exploratory Data Analysis

The data exploration stage is carried out to understand the characteristics and quality of the dataset before further analysis is performed. One important step in this exploration is analyzing the distribution of review lengths by calculating the average number of words per review. This analysis is necessary to identify variations

in text length that may affect the consistency of the analysis results. Reviews that exceed 20 words are then processed by segmenting sentences based on punctuation marks such as periods (.), question marks (?), and exclamation marks (!) as natural boundaries between sentences [27]. This process results in a total of 1,752 reviews, and this stage helps clarify the identification of aspects in the analysis [28].

In the next phase, language identification was performed for each review. The analysis results show the language distribution as follows: Indonesian (1,560 reviews), English (146 reviews), and Javanese and others (46 reviews). To complement the exploratory analysis, a word cloud visualization was created showing the frequency distribution of dominant words in the reviews (Figure 4).

Figure 4. (a) Wordcloud in Indonesian language, (b) Wordcloud in English language

Figure 4 presents a visualization of a word cloud that illustrates the distribution of dominant words in each language category. Figure (a) displays the word cloud in Indonesian, while Figure (b) shows the word cloud in English. The word clouds were generated from user reviews before undergoing preprocessing. This visualization facilitates the identification of word frequency patterns, which can support further analysis.

3.3 Preprocessing

The next stage is the data preprocessing stage, which is used to improve data quality before entering the analysis and modeling process. In this study, the preprocessing steps applied sequentially include cleaning, case folding, normalization, translation, synonym, stopword removal, and tokenization.

All reviews, both written in Indonesian and English, first go through the cleaning, case folding, and normalization processes. At the translation stage, all Indonesian reviews are then translated into English, resulting in a corpus that is entirely in English. This strategy was chosen to simplify the subsequent analysis process and allow the use of an English-based integrated processing approach. After the translation process is complete, all data is then processed through the synonym stages, stopword removal (based on the English list), and tokenization. Examples of data results that have gone through the preprocessing process can be seen in Table 1.

Table 1. Preprocessing text

Before preprocessing	After Preprocessing
Sampai sekarang akun tidak bisa aktif karena kode OTP tidak masuk email.	['account', 'cant', 'active', 'otp', 'code', 'not', 'included', 'email']
I'M REALLY TO LOVE “SUROBOYO BUS”, Because THE BEST	['love', 'bus', 'best']

3.4 Topic Modeling

The topic modeling stage in this study was carried out using the Latent Dirichlet Allocation (LDA) method to identify and group the main themes emerging from user reviews of the GOBIS Suroboyo application. LDA was chosen for its ability to uncover latent structures in text data in an unsupervised manner. The model was developed using the *gensim* library, with key parameters explicitly defined for reproducibility, including alpha=0.5, passes=5, chunksize=50, random_state=56, and per_word_topics=True. The optimal number of topics was determined based on a combined approach using coherence scores and stability score. The topic coherence was measured using the *c_v* metric, which considers word co-occurrence and semantic validity. Meanwhile, topic stability was calculated using Jaccard similarity between topic sets generated from different topic numbers (e.g., between models with *k* and *k+1* topics), to assess thematic consistency. The difference between coherence and stability scores was then analyzed, and the number of topics with the greatest difference was selected as the optimal number, as it provides the best balance between interpretability and result consistency. A visualization of the trends in coherence and stability scores across topic numbers is presented in Figure 5.

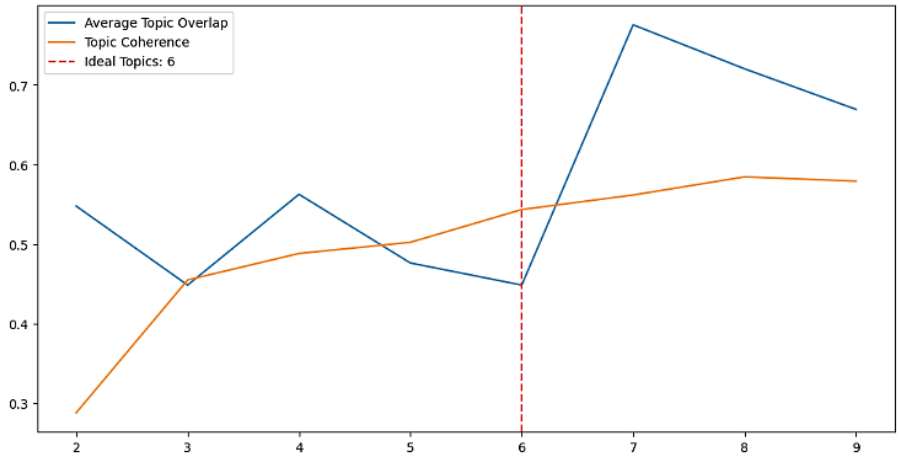

Figure 5. Best of topic modeling

The graph above shows that six topics is the most optimal number, indicated by a relatively high coherence score and a low degree of overlap between topics, which indicates good model stability. Therefore, the selection of the best number of topics is based on a combination of coherence scores and stability scores, as this approach can produce a balanced model in terms of interpretability and consistency. Based on Figure 5, six topics are established as the best number of topics, with the keywords and topic names of each topic presented in Table 2.

Table 2. Topic modeling

Topic	Keywords	Aspect Name
1	0.0459*”update” + 0.0227*”application” + 0.0186*”need” + 0.0126*”cool” + 0.0119*”development” + 0.0103*”gobis” + 0.0093*”clear” + 0.0091*”notification” + 0.0073*”pay” + 0.0073”citizens”	Application Features and Development
2	0.0415*”add” + 0.0218*”display” + 0.0141*”hope” + 0.0102*”admin” + 0.0094*”distance” + 0.0088*” search” + 0.0081*”suggestions” + 0.0079*”advice” + 0.0077*”mikrolet” + 0.0074*” innovation”	User Suggestions and Service Innovation
3	0.0300*”error” + 0.0260*”map” + 0.0255*”gps” + 0.0158*”information” + 0.0155*”not” + 0.0153*”come” + 0.0098*”cant” + 0.0095*”quite” + 0.0088*”feature” + 0.0080*”little”	Error and Location Accuracy
4	0.0619*”schedule” + 0.0487*”waiting” + 0.0179*”friendly” + 0.0157*”less” + 0.0156*”stop” + 0.0152*”estimated” + 0.0137*”bus” + 0.0125*”not” + 0.0121*”function” + 0.0110*”passengers”	Delay and Application Usability
5	0.0861*”good” + 0.0447*”application” + 0.0210*”arrive” + 0.0158*”using” + 0.0134*”hour” + 0.0105*”comfortable” + 0.0102*”improve” + 0.0092*”fleet” + 0.0089*”service” + 0.0074*”leave”	Comfort and Service Quality
6	0.0763*”bus” + 0.0530*”location” + 0.0428*”helpful” + 0.0406*”route” + 0.0377*”track” + 0.0289*”stop” + 0.0248*”user” + 0.0235*”not” + 0.0224*”application” + 0.0167*”transportation”	Route Tracking and Vehicle Information

3.5 Labeling

The labeling stages consist of two main processes, namely aspect labeling and sentiment labeling. For aspect labeling, it utilizes the dominant topics generated from the LDA topic model distribution. The topics with the highest probability are selected as aspect labels because they are considered to most represent the review content. The distribution of the number of aspect labels for each category can be seen in Table 3.

Table 3. Results of aspect labeling

Aspect	Frequency
Application Features and Development	145
User Suggestions and Service Innovation	126
Error and Location Accuracy	213
Delay and Application Usability	248
Comfort and Service Quality	249
Route Tracking and Vehicle Information	413

After the aspect labeling process is complete, the data is then further analyzed through sentiment labeling using a lexicon-based approach. This method combines lexical dictionaries from TextBlob and Pattern, and is supported by manual labeling to improve accuracy. After the three annotators complete the sentiment labeling, the next step is to calculate Fleiss' Kappa value. The obtained value is 0.4064, which indicates a moderate level of agreement among the annotators. The distribution of sentiment label counts is presented in the following table.

Table 4. Results of sentiment labeling

Sentiment	Frequency
Positive	792
Negative	602

Based on Table 4 above, it can be seen that positive sentiment has a frequency of 792, while negative sentiment is recorded at 602. This indicates that overall, positive sentiment is more dominant than negative sentiment in the analyzed data. To understand the distribution of sentiment in more detail, the image below will visualize the spread of positive and negative sentiment in each aspect.

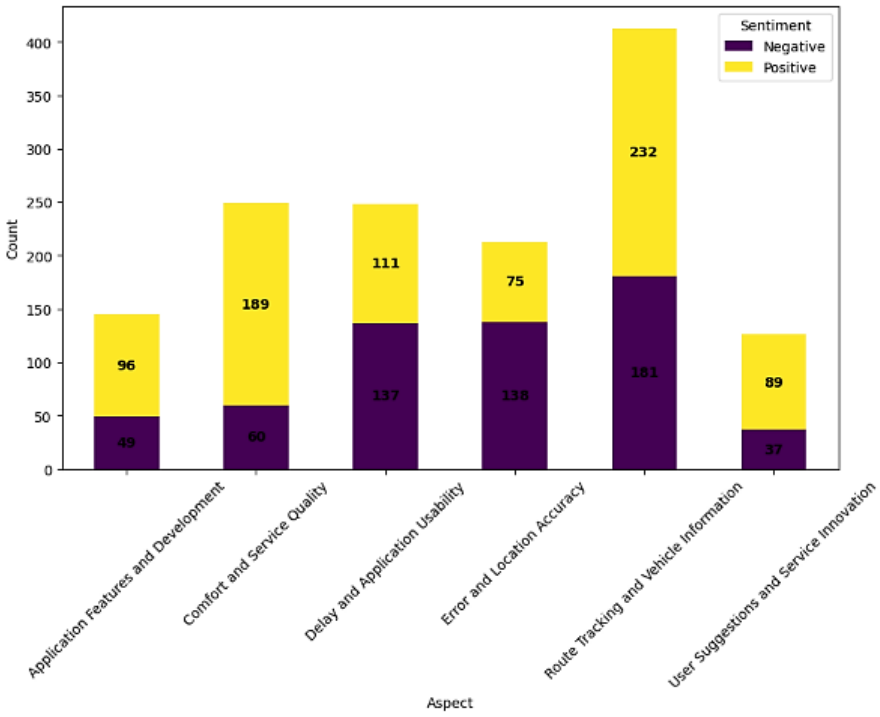

Figure 6. Bar Chart of Sentiment Distribution for Each Aspect

In the image above, Figure 6 shows the sentiment distribution from the labeling results indicating that each aspect has a different sentiment distribution. In the Application Features and Development aspect, there are 96 positive reviews and 49 negative reviews. The Comfort and Service Quality aspect records 189 positive reviews and 60 negative reviews. Meanwhile, the Delay and Application Usability aspect shows a contrasting trend, with 137 negative reviews and 111 positive reviews. A similar situation occurs in the Error and Location Accuracy aspect, which has 138 negative reviews and only 75 positive reviews. The Route Tracking and Vehicle Information aspect has the highest number of reviews, totaling 232 positive reviews and 181 negative reviews. Finally, the User Suggestions and Service Innovation aspect received 89 positive reviews and 37 negative reviews.

Table 5 below supports the findings above by displaying examples of aspect and sentiment labeling from the sample data. This table illustrates the relationship between dominant topics in reviews and the associated sentiments.

Table 5. Results of Aspect and Sentiment Labeling

Review	Aspect	Sentiment
Aplikasinya lemot bgt, gak bisa muncul bisnya. Lokasi juga gak sesuai sama titik di mapnya.	Route Tracking and Vehicle Information	Negative
I think this app needed an update because it have so many bugs in the app, example: I can’t press the log out button and log in button and I can’t see the bus location while I can’t log in.	User Suggestions and Service Innovation	Negative

Table 5 above presents an example of the labeling results of aspects and dominant sentiments in the dataset, which is determined based on the principle of majority voting, that is, if a label appears at least twice out of three annotators. Through this approach, each review receives a final label based on the majority agreement among the evaluators.

3.6 Data Splitting

At this stage, the process of dividing the data into two main parts is conducted, namely data modeling and data validation. The splitting technique used is the hold-out method with a ratio of 90:10 of the total available data. As much as 90% of the data is allocated for data modeling that will be used in the development and training process of the model. Meanwhile, the remaining 10% is used as final validation data to test and evaluate the performance of the system developed during the modeling stage. With the final results of data modeling consisting of 1254 rows, and validation data of 140 rows.

3.7 Model Development

The next stage involves building a classification model using 90% of the pre-split dataset. This modeling process involves several steps. First, the modeling dataset is split using various evaluation scenarios: hold-out method with 90:10, 80:20, and 70:30 proportions, and cross-validation using 5-fold and 10-fold splits. In each scenario, word weighting is performed using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. Next, text classification for sentiment analysis is performed using the Support Vector Machine (SVM) algorithm, based on sentiment labels obtained through a hybrid approach combining lexicon-based and manual annotations. Classification is performed using the LinearSVC implementation of the SVM algorithm. The linear kernel is chosen due to its effectiveness and computational efficiency for high-dimensional text data. The regularization parameter is set to its default value (C=1.0) and random_state=42 is set to ensure reproducibility. The final step involves evaluating the model using several metrics, including precision, recall, F1 score, and accuracy. The results of applying SVM with various data sharing scenarios produce a classification report which can be seen in the following table.

Table 6. Data Splitting Scenario

Data Splitting		Accuracy	Precision	Recall	F1-Score
Hold-out	90:10	0.7031	0.6990	0.7004	0.6995
	80:20	0.7312	0.7274	0.7242	0.7254
	70:30	0.7150	0.7103	0.7109	0.7106
Cross-Val	5-fold	0.7297	0.7256	0.7221	0.7234
	10-fold	**0.7416**	**0.7376**	**0.7354**	**0.7363**

Based on Table 6 above, it can be concluded that the 10-fold Cross-Validation scenario shows the best performance compared to other scenarios. This model achieved the highest value on all main evaluation metrics, with details as follows: accuracy 74.16%, precision 73.76%, recall 73.54%, and F1-score 73.63%.

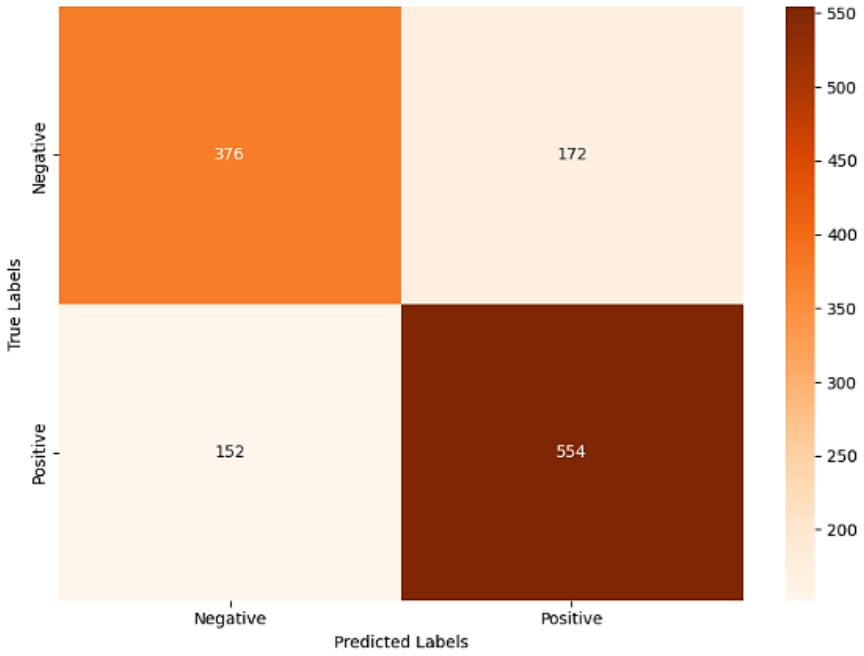

Figure 7. Confusion matrix

Based on Figure 7, which displays the confusion matrix of the SVM model evaluation results, it can be seen that this model has quite good predictive performance. The main achievement is evident in the dominant number of accurate predictions, with 376 true negatives (correct negative predictions) and 554 true positives (correct positive predictions). Although there are 172 false positives (negatives predicted as positives) and 152 false negatives (positives predicted as negatives), the overall accuracy of the model is quite good, with a dominance of correct predictions in both classes. The relatively balanced error rate between false positives and false negatives indicates that the model does not have significant bias toward either class, and has adequate generalization capability for this sentiment classification.

3.8 Model Validation

The purpose of this stage is to test whether the aspect and sentiment analysis features function well overall. The validation process uses a dedicated dataset that has been separated since the early stages (a hold-out validation set of 10%) and is completely uninvolved in the model training process. This characteristic makes the validation dataset ideal for measuring the model's generalization ability on new data that has never been processed before.

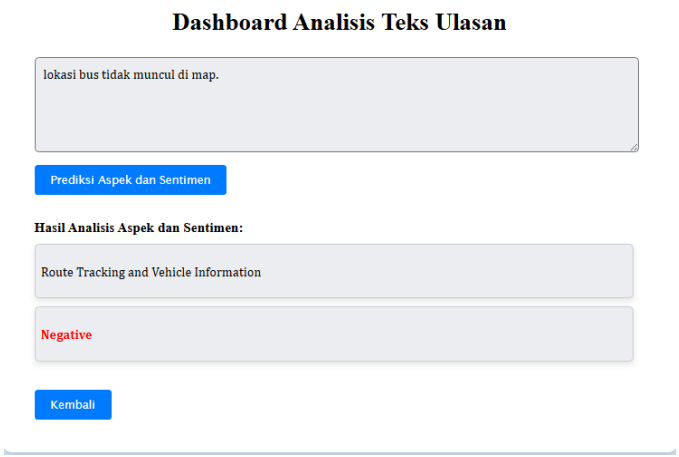

Figure 8. Example reviews in Indonesia language

For illustration, Figure 8 displays evidence of the model's performance through real examples of processing reviews in Indonesian, which also validates the model's effectiveness in handling Indonesian language text data.

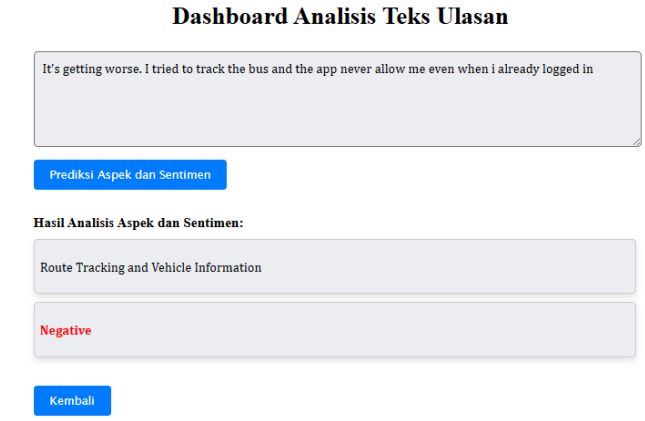

Figure 9. Example reviews in English language

In addition, Figure 9 demonstrates the same capability for English language reviews. This result proves the reliability of the model in handling English language text data.

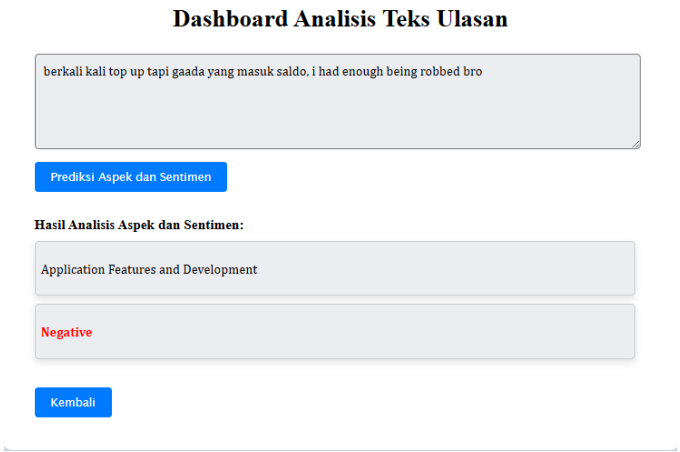

Figure 10. Example reviews in bilingual manner (Indonesian-English)

This model also successfully processes reviews in a bilingual manner (Indonesian-English) as shown in Figure 10, proving its capability in handling multilingual data as well as the validity of its application in real-world scenarios.

3.9 Limitations of the Results

Although this study has succeeded in building a fairly good model and is able to predict multilingual reviews (Indonesian and English), this study still has several limitations that need to be considered. First, the data used is limited to user reviews from one platform, namely the Google Play Store with a total of 1,168 reviews. This can limit the representativeness of the data to the entire GOBIS user population. Second, although LDA is effective in extracting latent aspects, this method still has weaknesses, such as producing overlapping topics or unclear interpretations. Third, the sentiment classification model built using the SVM algorithm showed quite good performance, but not optimal, with an accuracy of 74.16%. Finally, the less than optimal accuracy resulted in some reviews being mispredicted both in terms of aspects and sentiment polarity.

3.10 Comparison with Prior Work

This study contributes significantly to the development of Aspect-Based Sentiment Analysis (ABSA) methods in the context of community-based public transportation services, particularly focusing on the GOBIS Suroboyo application. In this regard, the results are compared with three relevant prior studies. First, the study by Airin et al. [7] applied ABSA to e-hailing service reviews in Malaysia using the LDA algorithm for aspect extraction, and evaluated the quality of the generated topics quantitatively using perplexity and coherence score. Their sentiment classification model demonstrated reasonably strong performance across standard evaluation metrics such as accuracy, precision, recall, and F1-score, and was tested on reviews written in two official languages, namely English and Malay. However, their approach did not fully address the complexity of code-mixed multilingual reviews and lacked a fully integrated classification system that spans from preprocessing to final classification. In contrast, the present study overcomes these limitations by developing a comprehensive linguistic pipeline, which includes slang normalization, automatic translation, and synonym substitution, and incorporates a stability score for topic evaluation an enhancement not employed in the previous study.

Second, the study conducted by Mustakim and Priyanta [8] analyzed user reviews of the KAI Access application using a supervised approach, implementing both Naïve Bayes and SVM algorithms. In that study, aspects were determined manually, and the data consisted solely of Indonesian-language reviews, without consideration of language variation or an unsupervised method for aspect extraction. Compared to that, the current research offers a more flexible and data-driven approach, utilizing LDA for automatic aspect extraction and SVM for sentiment classification, enabling effective handling of reviews written in diverse and mixed-language formats.

Third, the study by Andono et al. [9] combined LDA with semantic similarity and the deep learning model BERT to analyze hotel reviews. Their classification model achieved very high performance, with F1-scores 0.97. Although their results outperformed those of this study in terms of evaluation metrics, this study yielded a maximum F1-score of 0.7363 using 10-fold cross-validation, their work was limited to the hotel review domain and did not offer a fully integrated end-to-end ABSA pipeline. Furthermore, multilingual or code-mixed review challenges were not the main focus of their study.

In summary, although the model performance metrics in this research are relatively moderate, its main strengths lie in the system's ability to handle code-mixed multilingual reviews, the use of more objective and quantitative topic evaluation methods, and the development of a comprehensive aspect and sentiment classification pipeline. As such, this study is expected to serve as a solid foundation for advancing public opinion analysis systems tailored to local public transportation services in Indonesia.

4. CONCLUSION

Based on the research results, aspect-based sentiment analysis on the GOBIS Suroboyo application successfully identified six dominant aspects in user reviews, namely Application Features and Development, User Suggestions and Service Innovation, Error and Location Accuracy, Delay and Application Usability, Comfort and Service Quality, as well as Route Tracking and Vehicle Information. The Support Vector Machine (SVM) based model with 10-fold cross-validation achieved consistent performance in classifying multilingual reviews, demonstrated by balanced evaluation metrics of accuracy (74.16%), precision (73.76%), recall (73.54%), and F1-score (73.63%). These results indicate that the model is not only able to classify positive and negative sentiments, but is also able to face linguistic challenges that arise in multilingual data. However, this study still has several limitations. One of them is that there are still a number of predictions that are less accurate, both in terms of aspects and sentiment. This shows that the model still needs to be improved, especially in terms of accuracy and precision in adjusting aspects to the context of the sentence. However, the findings of this study have the potential to make a real contribution to the development of public services, especially in the field of smart transportation. By knowing the aspects that are most often appreciated and complained about by users, related agencies such as the Surabaya City Transportation Department can use the results of this analysis as a basis for evaluating and improving GOBIS services in a more targeted manner. In addition, the information obtained can also be used as a consideration in formulating data-based public policies, in order to increase public satisfaction with public transportation services and support smart city initiatives that are oriented towards citizen needs. This analytical approach can also be replicated for other public service applications as part of the government's digital service quality improvement strategy.

REFERENCE

[1] B. P. S. K. Surabaya, "Kota Surabaya Dalam Angka 2024 [Surabaya City in Figures 2024]," 2024. (In Indonesian)

[2] Z. A. Haqie, R. E. Nadiah, and O. P. Ariyani, "Inovasi Pelayanan Publik Suroboyo Bis di Kota Surabaya [Innovation of Public Services Suroboyo Bis in Surabaya City]," *JPSI (Journal Public Sect. Innov.*, vol. 5, no. 1, p. 23, Dec. 2020. https://dx.doi.org/10.26740/jpsi.v5n1.p23-30 (In Indonesian)

[3] R. A. Firmansyah and K. H. Putra, "Analisis Tingkat Kepuasan Pengguna Transportasi Umum 'Suroboyo Bus' Rute Halte Rajawali-Terminal Purabaya Dengan Metode Importance Performance Analysis (Ipa) [Analysis of User Satisfaction Level of Public Transportation 'Suroboyo Bus' Route Rajawali Bus Stop-Purabaya Terminal Using Importance Performance Analysis (IPA) Method]," *Semin. Teknol. Perencanaan, Perancangan, Lingkungan, dan Infrastruktur FTSP ITATS*, pp. 1–6, 2019, [Online]. Available: https://ejurnal.itats.ac.id/stepplan/article/view/711/612. (In Indonesian)

[4] S. Roiqoh, B. Zaman, and K. Kartono, "Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Mobile JKN dengan Lexicon Based dan Naïve Bayes [Aspect-Based Sentiment Analysis of JKN Mobile Application Reviews with Lexicon Based and Naïve Bayes]," *J. Media Inform. Budidarma*, vol. 7, no. 3, p. 1582, 2023. https://dx.doi.org/10.30865/mib.v7i3.6194 (In Indonesian)

[5] R. A. Rahman, V. H. Pranatawijaya, and N. N. K. Sari, "Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Gojek [Aspect-Based Sentiment Analysis on Gojek App Reviews]," *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 4, no. 1, pp. 70–82, Jun. 2024. https://dx.doi.org/10.24002/konstelasi.v4i1.8922 (In Indonesian)

[6] Y. Kustiyahningsih and Y. Permana, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) Untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink [Use of Latent Dirichlet Allocation (LDA) and Support-Vector Machine (SVM) to Analyze Sentiment Based on Aspects in EdLink Application Reviews]," *Teknika*, vol. 13, no. 1, pp. 127–136, 2024. https://dx.doi.org/10.34148/teknika.v13i1.746 (In Indonesian)

[7] K. A. F. A. Samah *et al.*, "Unveiling perceptions: aspect-based sentiment analysis of Malaysia's e-hailing reviews," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 37, no. 3, p. 2077, Mar. 2025. https://dx.doi.org/10.11591/ijeecs.v37.i3.pp2077-2086

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

[9] P. N. Andono, Sunardi, R. A. Nugroho, and B. Harjo, "Aspect-Based Sentiment Analysis for Hotel Review Using LDA, Semantic Similarity, and BERT," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 5, pp. 232–243, 2022. https://dx.doi.org/10.22266/ijies2022.1031.21

[10] A. Hadi, M. Qamal, and Y. Afrillia, "Comparison of Random Forest Algorithm Classifier and Naïve Bayes Algorithm in Whatsapp Message Type Classification," *J. Renew. Energy, Electr. Comput. Eng.*, vol. 5, no. 1, pp. 9–17, Mar. 2025. https://dx.doi.org/10.29103/jreece.v5i1.21227

[11] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "The Crowdsourcing Method to Normalize 'Bahasa Alay', a Case of Indonesian Corpus," in *2020 Fifth International Conference on Informatics*

and Computing (ICIC), IEEE, Nov. 2020, pp. 1–5. https://dx.doi.org/10.1109/ICIC50835.2020.9288534

[12] H. Tuhuteru, L. P. Refialy, M. Laturake, and S. G. Pattirane, "Tourist Perceptions Through Sentiment Analysis to Support Tourism Development in Maluku Province," *J. Appl. Informatics Comput.*, vol. 8, no. 1, p. 119, 2024. https://dx.doi.org/10.30871/jaic.v8i1.6989

[13] A. Gerliandeva, Y. H. Chrisnanto, and H. Ashaury, "Optimasi Klasifikasi Sentimen pada Komentar Online menggunakan Multinomial Naïve Bayes dan Ekstraksi Fitur TF-IDF serta N-grams Optimization of Sentiment Classification on Online Comments using Multinomial Naïve Bayes and TF-IDF Feature Extraction and N-g [Optimization of Sentiment Classification on Online Comments using Multinomial Naïve Bayes and TF-IDF Feature Extraction and N-grams]," *Pekommas*, vol. 9, no. X, pp. 259–272, 2024. https://dx.doi.org/10.56873/jpkm.v9i2.5585 (In Indonesian)

[14] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, p. e0232525, May 2020. https://dx.doi.org/10.1371/journal.pone.0232525

[15] Z. Sitorus, M. Saputra, S. N. Sofyan, and Susilawati, "Sentiment Analysis of Indonesian Community Towards Electric Motorcycles on Twitter Using Orange Data Mining," *INFOTECH J.*, vol. 10, no. 1, pp. 108–113, 2024. https://dx.doi.org/10.31949/infotech.v10i1.9374

[16] E. S. Negara and D. Triadi, "Topic modeling using latent dirichlet allocation (LDA) on twitter data with Indonesia keyword," *Bull. Soc. Informatics Theory Appl.*, vol. 5, no. 2, pp. 124–132, 2022. https://dx.doi.org/10.31763/businta.v5i2.455

[17] S. Dami and R. Alimardani, "An Aspect-Level Sentiment Analysis Based on LDA Topic Modeling," *J. Inf. Syst. Telecommun.*, vol. 12, no. 2, pp. 117–126, 2024.

[18] N. Kalaitzidis, "What is the best way to obtain the optimal number of topics for a LDA-Model using Gensim?" [Online]. Available: https://stackoverflow.com/questions/32313062/what-is-the-best-way-to-obtain-the-optimal-number-of-topics-for-a-lda-model-usin

[19] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek [Comparison of Naïve Bayes, SVM, and k-NN for Aspect-Based Gadget Sentiment Analysis]," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 6, pp. 1120–1126, 2021. https://dx.doi.org/10.29207/resti.v5i6.3588 (In Indonesian)

[20] M. Venugopalan and D. Gupta, "An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis," *Knowledge-Based Syst.*, vol. 246, p. 108668, 2022. https://dx.doi.org/10.1016/j.knosys.2022.108668

[21] B. Ozyurt and M. A. Akcayol, "A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA," *Expert Syst. Appl.*, vol. 168, p. 114231, 2021. https://dx.doi.org/10.1016/j.eswa.2020.114231

[22] Z. A. Khan *et al.*, "Developing Lexicons for Enhanced Sentiment Analysis in Software Engineering: An Innovative Multilingual Approach for Social Media Reviews," *Comput. Mater. Contin.*, vol. 79, no. 2, pp. 2771–2793, 2024. https://dx.doi.org/10.32604/cmc.2024.046897

[23] A. Mahmoudi, D. Jemielniak, and L. Ciechanowski, "Assessing Accuracy: A Study of Lexicon and Rule-Based Packages in R and Python for Sentiment Analysis," *IEEE Access*, vol. 12, no. November 2023, pp. 20169–20180, 2024. https://dx.doi.org/10.1109/ACCESS.2024.3353692

[24] R. U. Fahmi, A. A. Arifiyanti, and T. L. I. Sugata, "Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Midi Kriing Menggunakan Support Vector Machine (SVM) [Aspect-Based Sentiment Analysis on Midi Kriing Application Review Using Support Vector Machine (SVM)]," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 3, pp. 4831–4839, May 2025. https://dx.doi.org/10.36040/jati.v9i3.13783 (In Indonesian)

[25] D. I. Sumantiawan, J. E. Suseno, and W. A. Syafei, "Sentiment Analysis of Customer Reviews Using Support Vector Machine and Smote-Tomek Links For Identify Customer Satisfaction," *J. Sist. Info. Bisnis*, vol. 13, no. 1, pp. 1–9, 2023. https://dx.doi.org/10.21456/vol13iss1pp1-9

[26] D. Rahmawati, E. D. Wahyuni, and D. S. Y. Kartika, "Analisis Sentimen Berbasis Aspek pada Respons Survei Open-Ended Menggunakan LDA, dan SVM [Aspect-Based Sentiment Analysis on Open-Ended Survey Responses Using LDA, and SVM]," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 3, pp. 4628–4634, May 2025. https://dx.doi.org/10.36040/jati.v9i3.13723 (In Indonesian)

[27] S. Raharjo, R. Wardoyo, and A. E. Putra, "Rule based sentence segmentation of Indonesian language," *J. Eng. Appl. Sci.*, vol. 13, no. 21, pp. 8986–8992, 2018. https://dx.doi.org/10.3923/jeasci.2018.8986-8992

[28] Y. Seo, S. Song, R. Heo, J. Kim, and D. Lee, "Make Compound Sentences Simple to Analyze: Learning to Split Sentences for Aspect-based Sentiment Analysis," *EMNLP 2024 - 2024 Conf. Empir. Methods Nat. Lang. Process. Find. EMNLP 2024*, pp. 11171–11184, 2024.