JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 8, No. 1, June 2025, pp. 97~111
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: 10.36378/jtos.v8i1.4297

Application of Ensemble Machine Learning Methods for Aspect-Based Sentiment Analysis on User Reviews of the Wondr by BNI App

Rendi Hardiartama[1], Amalia Anjani Arifiyanti[2], Seftin Fitri Ana Wati[3]
[1,2,3]Department of Computer Science, University of Pembangunan Nasional "Veteran" Jawa Timur , Indonesia

Article Info

Article history:

Received 05 05, 2025
Revised 05 15, 2025
Accepted 05 20, 2025

Keywords:

ABSA
Stacking Ensemble Learning
Scraping
LIME
Wondr By BNI

ABSTRACT

This study analyzes user perceptions of the Wondr by BNI app using Aspect-Based Sentiment Analysis (ABSA) and a stacking ensemble learning approach on user reviews. Data were collected from the Google Play Store and App Store through scraping, then processed and labeled. The study involves two classification stages: aspect identification and sentiment classification for each aspect. The stacking ensemble model without resampling showed the best performance, with F1-scores of 99.4% for UI (User Interface), 99.3% for Authentication, and 99% for Transaction. For sentiment classification, F1-scores reached 82.2% User Interface (UI), 87.8% (Authentication), and 92.4% (Transaction). The use of LIME (Local Interpretable Model-Agnostic Explanations) improved model interpretability by highlighting keywords influencing the classification results. The final output of this research is a website capable of performing aspect-based sentiment classification.

Corresponding Author:

Rendi Hardiartama
Department of Computer Science
University of Pembangunan Nasional "Veteran" Jawa Timur
Jawa Timur, Indonesia
Email: rendirenhard@mail.com

1. Introduction

The development of information and communication technology has driven digital transformation in the banking sector, one of which is through mobile banking [1]. This service enables users to conduct financial transactions such as transfers, payments, and balance checks directly via smartphones [2].

In Indonesia, mobile banking usage has shown significant growth, with several major banks reporting a surge in the number of users. PT Bank Rakyat Indonesia (Persero) Tbk (BRI), through the BRImo app, reached 33.5 million users, a 30.3% increase compared to Q1 2023. PT Bank Central Asia Tbk (BCA) reported 30.8 million mobile banking users, growing by 9%. PT Bank Mandiri (Persero) Tbk, with the Super App Livin' by Mandiri, saw a 39% growth to 24 million users in Q1 2024. However, PT Bank Negara Indonesia (Persero) Tbk (BNI), despite a growth of 18.5% to 16.9 million users in Q1 2024, still lags behind its competitors [3].

To strengthen its position in digital banking, BNI launched the Wondr by BNI application in July 2024 as its latest mobile banking service, available for free on the Google Play Store [4][5]. As of December 31, 2024, the app recorded over 5 million downloads with a rating of 3.8, lower than BNI Mobile Banking which has a rating of 4.5. The lower rating indicates that user satisfaction is not yet optimal [6], making user reviews important for gaining deeper insights into the app's performance [7]. Reviews not only help other users but also provide valuable feedback for developers [8]. The large volume of available reviews makes sentiment analysis necessary to filter and extract meaningful insights. Therefore, an Aspect-Based Sentiment Analysis (ABSA) approach is needed, which can identify specific aspects in reviews and determine the sentiment related to each aspect [9][10].

ABSA research has employed various machine learning algorithms such as Naive Bayes, SVM, and Decision Tree [11]. However, class imbalance challenges often cause models to be biased toward the majority class [12][3]. To overcome this, ensemble learning methods are used, particularly the stacking technique, which combines several base learners and one meta-learner to improve accuracy [14][15]. Stacking is chosen because it has been proven superior to other ensemble methods in terms of accuracy, prediction stability, and resistance to overfitting [16].

Nonetheless, this method faces challenges related to interpretability, as prediction results are often considered a "black box" or the model does not provide explanations for its decisions [18]. To enhance transparency and trust in the model, an interpretative approach is required. One such method is LIME (Local Interpretable Model-agnostic Explanations), which provides local explanations for prediction results without depending on any specific model type [19].

Two studies applied stacking ensemble techniques to improve sentiment analysis. The first [17] proposed a framework for Aspect-Based Sentiment Analysis (ABSA) using various models (Logistic Regression, Random Forest, XGBoost, MLP, and SVM), with Logistic Regression as a meta-model. It significantly improved F1 scores for ACD (↑22.9–28.4%) and accuracy for ACP (↑9.3–13.2%). The second study [19] used a similar approach on online loan app reviews, combining Naïve Bayes, Random Forest, and SVM, achieving 87.05% accuracy. To enhance interpretability, LIME was applied to highlight key words influencing predictions, increasing transparency and trust in the model.

This research aims to analyze aspects and sentiments in reviews of the Wondr by BNI application collected through scraping from the Play Store and App Store. The analysis process involves two classification stages: aspect identification and sentiment classification for each aspect, as well as model interpretation. The final output is a review analysis website expected to provide strategic insights for developing an application that better meets user needs.

2. Research Method

2.1. Research Flow

The research flow outlines the stages of the study from beginning to end. The research flow is presented in Figure 1 below.

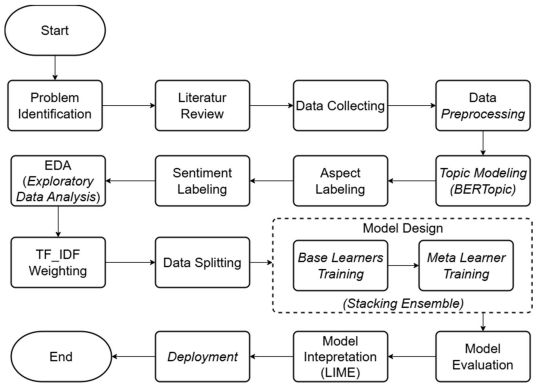

Figure 1. Research Flow

2.2. Problem Identification

Problem identification is the initial stage of research aimed at determining the issue to be studied, beginning with the observation of a phenomenon, then formulated specifically as the foundation of the research.

2.3. Literature Review

The second step is conducting a literature review to gain a deeper understanding of the problem, exploring concepts such as ABSA and Stacking Ensemble. References are taken from journals, articles, and scientific works to strengthen the theoretical foundation and support data analysis.

2.4. Data Collection

Reviews of the Wondr by BNI application were collected through web scraping using google-play-scraper and app_store_scraper. The reviews were filtered from July 31 to December 31, 2024.

2.5. Data Preprocessing

Data preprocessing is a crucial step in text analysis to prepare and clean the data for analysis [20].

1. Data Cleaning: Removing irrelevant elements or noise such as punctuation, numbers, and special characters.
2. Case Folding: Converting all letters in the text to lowercase to avoid discrepancies between identically spelled words with different capitalizations.
3. Normalization: Standardizing word or phrase forms, such as replacing abbreviations or acronyms with their full forms.
4. Stopword Removal: Removing frequently occurring but unimportant words like “and”, “yes” or “is.”
5. Tokenizing: Splitting text into smaller units (tokens), such as words or phrases.
6. Stemming, Returning a word to its root form

2.6. Topic Modeling

Topic modeling was conducted using BERTopic, which combines BERT-based text representations, dimensionality reduction via UMAP (Uniform Manifold Approximation and Projection), clustering with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), and keyword extraction using C-TF-IDF (Clustered Term Frequency-Inverse Document Frequency). This process produces structured and interpretable topics based on relevant keywords in each cluster [21].

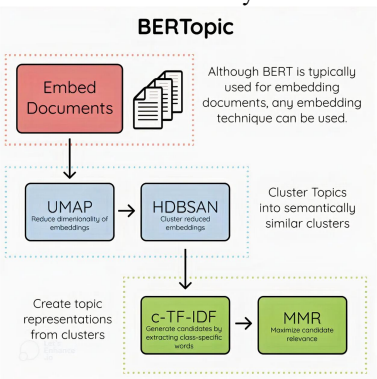

Figure 2. Flow BERTopic [22]

1. Embed Documents: Text documents are transformed into vector representations using embedding models like BERT.
2. Dimensionality Reduction with UMAP: UMAP is used to reduce the dimensions of the embedding while preserving the semantic structure of the data.
3. Clustering with HDBSCAN: The data is clustered using HDBSCAN to identify topic clusters based on data density in the embedding space.
4. Topic Representation with c-TF-IDF: c-TF-IDF is used to extract the most relevant words in each cluster, creating meaningful topic descriptions.
5. MMR (Maximal Marginal Relevance): MMR is used to select the most relevant and non-redundant key words, resulting in more unique topics.

2.7. Aspect Labeling

This study employs a Cosine Similarity approach, which evaluates the similarity between words for automatic aspect labeling. Aspect and document representations are converted into vectors using Word2Vec,

and their similarity scores are then calculated. If the cosine similarity score is ≥ 0.8, the document is labeled with the aspect that has the highest similarity; if the score is < 0.8, the document is not labeled. This approach assesses the semantic similarity of text and is effective for analyzing reviews or opinions [23]. The mathematical formula used to calculate Cosine Similarity is as follows [28] :

$$\text{Cosine Similarity} = \cos(\theta) = \frac{A \cdot B}{||A||\,||B||} \quad (1)$$

Formula description :

$A \cdot B$, is the dot product of vectors A and B, calculated as $\boldsymbol{\Sigma}_{k=1}^{n} x_k\, y_k$ (2)

$||A||$ is the magnitude (length) of vector A, calculated as $\sqrt{\boldsymbol{\Sigma}_{k=1}^{n} x_k^2}$ (3)

$||B||$ is the magnitude (length) of vector B, calculated as $\sqrt{\boldsymbol{\Sigma}_{k=1}^{n} y_k^2}$ (4)

x_k is the weight of term k in vector x_k

y_k is the weight of term k in vector y_k

n is the number of dimensions (terms) in the vectors

Table 1. Criteria Correlation Pearson [23]

Nilai r	Criteria Correlation
0	No correlation
0 - 0.5	Weak correlation
0.5 - 0.8	Moderate correlation
0.8 - 1	Strong correlation
1	Perfect correlation (2 identical words)

2.8. Sentiment Labeling

Sentiment labeling is performed automatically using a lexical window of context approach with a parameter $k = 3$ to capture the local context around aspects. This approach assumes that words describing opinions about an aspect are usually located near that aspect [24]. Below is an explanation of the symbols used in the formula:

$$\text{Context}_{lex}(w_x) = \begin{cases} [w_{x-k} : w_{x+k}] & \text{if } n-k > x > k \\ [w_1 : w_{x+k}] & \text{if } x < k \\ [w_{x-k} : w_n] & \text{if } x + k > n \end{cases} \quad (5)$$

w: Word representing the aspect word in the text or word sequence.
k: Context window size, which is the number of words before or after the word w_x that we want to take as context. This is usually referred to as "window size" in natural language processing.
n: The total length of the text or word sequence. This refers to the number of words in the document or corpus.
x: The position of the word w_x in the word sequence (for example, if w_x is the third word in the sequence, then x = 3).

For example, in the sentence: "The food was amazing but the service could have been better," If the aspect being analyzed is "food," then the lexical window with k=3 will include the words around the aspect, i.e., "The," "was amazing but."

Then, this approach is implemented using three sentiment labeling techniques: the InSet Indonesian, the Indonesian RoBERTa Base Sentiment Classifier, and the Indonesian RoBERTa Base IndoLEM Sentiment Classifier. The final sentiment label is determined by taking the mode of the three results, followed by manual evaluation to assess the effectiveness of the labeling.

2.9. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) aims to understand the structure and patterns in the review data, including the distribution of aspects and sentiments. Visualizations such as charts and word clouds are used to present dominant keywords and sentiment distribution per aspect.

2.10. TF-IDF weighting

TF-IDF weighting is used to transform text into numerical vectors representing the importance of words within a document. Using TfidfVectorizer from sklearn, this method gives higher weights to rare and relevant words while reducing the weight of common words.

2.11. Data Splitting

Data is split using the hold-out method, dividing the dataset into training data (80%) and test data (20%). The training data is used to train the model, while the test data is used to evaluate model performance. Oversampling is also applied using SMOTE.

2.12. Model Design

In the model design phase, a stacking ensemble approach is used to improve prediction accuracy by combining the strengths of multiple machine learning algorithms. Stacking is an Ensemble Learning technique that utilizes multiple base models trained in parallel and independently. The predictions from each base model are then combined using a meta-learning algorithm to produce the final output [13].

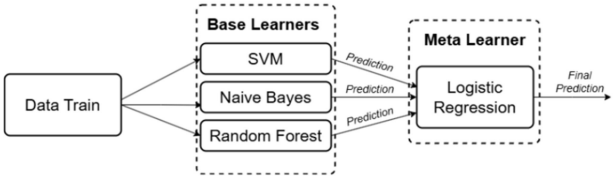

Figure 3. Architecture Stacking Ensemble

In the base learner, the algorithms used are SVM (Support Vector Machine), Naive Bayes, and Random Forest, while in the meta learner, the algorithm used is Logistic Regression.

2.13. Model Evaluation

The model evaluation stage aims to measure the performance of both aspect and sentiment models using a confusion matrix. The evaluation includes calculating accuracy, precision, recall, and F1 score [25].

1. Accuracy: Measures how well the model correctly predicts outcomes.
2. Precision: Assesses the accuracy of positive predictions, indicating how many of the predicted positives are actually correct.
3. Recall: Measures how effectively the model identifies all actual positive cases.
4. F1 Score: A metric that combines precision and recall to provide an overall view of model performance, especially useful in cases of class imbalance.

2.14. Model Interpretation

This study uses LIME (Local Interpretable Model-Agnostic Explanations) to assist in model interpretation. LIME is a model-agnostic method that provides local explanations for predictions made by black-box models. It works by generating explanations after the model has been built, enabling understanding of the factors influencing the prediction results [26].

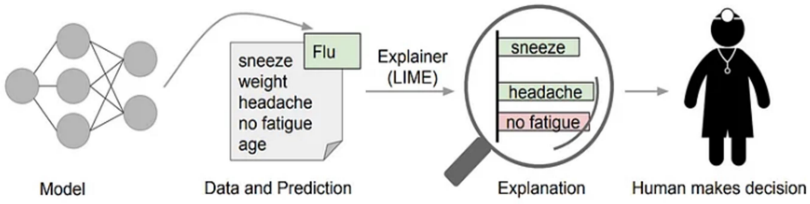

Figure 4. Explanation of individual predictions [27].

A model predicts that a patient has the flu, and LIME highlights the symptoms from the patient's medical history that support this prediction. Symptoms like sneezing and headache are considered supportive

of the flu diagnosis, while the absence of fatigue serves as evidence against it. With this information, a doctor can make a more informed decision about the reliability of the model's prediction .

2.15. Deployment

In the deployment stage, a Flask-based website is implemented to apply the best-performing model. Key features include an interactive dashboard with visualizations of key information, aspect-based sentiment analysis through direct text input, and batch analysis capability using CSV or XLSX files.

3. Result and Discussion

3.1. Data Collection

The review data obtained from the Play Store and App Store was collected within the period between July 31, 2024, and December 31, 2024, totaling 20,620 reviews. Of this total, 16,476 reviews came from the Play Store, while 4,144 reviews were sourced from the App Store.

3.2. Data Preprocessing

At this stage, preparation is carried out before modeling to ensure the data is cleaner, more structured, and ready for use.

Table 2. Data Preprocessing

	Before	After
Data Cleaning	Baru coba dan bni emang ok bisa bertransaki apapun enggk ribet👍👍👍	Baru coba dan bni emang ok bisa bertransaki apapun enggk ribet
Case Folding	Baru coba dan bni emang ok bisa bertransaki apapun enggk ribet	baru coba dan bni emang ok bisa bertransaki apapun enggk ribet
Normalization	baru coba dan bni emang ok bisa bertransaki apapun enggk ribet	baru coba dan bni memang ok bisa bertransaksi apapun enggak ribet
Stopword Removal	baru coba dan bni memang ok bisa bertransaksi apapun enggak ribet	coba bni bertransaksi apapun enggak ribet
Tokenizing	coba bni ok bertransaki apapun enggak ribet	[‘coba’,’bni’,’ok’,’bertransaksi’,’apapun’,’enggak’,’ribet’]
Stemming	[‘coba’,’bni’,’ok’,’bertransaksi’,’apapun’,’enggak’,’ribet’]	[‘coba’,’bni’,’ok’,’’transaksi’,’apa’,’enggak’,’ribet’]

After performing data cleaning, including the removal of duplicates and one-word reviews, the remaining number of reviews is 18,219.

3.3. Topic Modeling Results Using BERTopic

The topic modeling process using BERTopic yielded three distinct topics, which can be represented as follows:

1. Topic 1: UI (User Interface)

```
model.get_topic(0)

[('mudah', 0.5904114852763012),
 ('tampilan', 0.5806176706068079),
 ('cepat', 0.5355401850958286),
 ('sangat', 0.5052076491304482),
 ('simpel', 0.4971537622534343),
 ('transaksi', 0.47497167501518894),
 ('mantap', 0.4704782594236069),
 ('menarik', 0.464871009819693),
 ('bagus', 0.46368540360290667),
 ('keren', 0.46072731682428636)]
```

Figure 5. Representation of Topic 1

This topic represents the User Interface (UI) aspect, as it includes words such as *tampilan* (interface), *simpel* (simple), *menarik* (attractive), *bagus* (good), and *keren* (cool), which reflect users' evaluations of the application's design and visual appeal.

2. Authentication

```
model.get_topic(1)

[('verifikasi', 0.8065197824113529),
 ('login', 0.7072272119246096),
 ('email', 0.6382380507544781),
 ('error', 0.6154411723895635),
 ('gagal', 0.5677338973396876),
 ('update', 0.5439501684375535),
 ('wajah', 0.5381834921021683),
 ('otp', 0.49940155928461283),
 ('kode', 0.4869003618606809),
 ('susah', 0.45708097321540947)]
```

Figure 6. Representation of Topic 2

This topic represents the authentication aspect, as it contains words such as *verifikasi* (verification), *login*, *email*, *otp*, *kode* (code), and *wajah* (face), which describe the processes and methods involved in user authentication when logging into the application.

3. Transaction

```
model.get_topic(2)

[('mobile', 0.6489180392190066),
 ('transfer', 0.621943786188827),
 ('tunai', 0.5957571774421296),
 ('tarik', 0.5471953284114953),
 ('kartu', 0.5344397634875032),
 ('saldo', 0.531263873270897),
 ('tanpa', 0.5202421136333084),
 ('fitur', 0.49336700995458627),
 ('banking', 0.49069513107481766),
 ('transaksi', 0.46746209897476615)]
```

Figure 7. Representation of Topic 3

This topic represents the transaction aspect, as it includes words such as *transfer*, *tunai* (cash), *tarik* (withdraw), *saldo* (balance), and *transaksi* (transaction), describing users' financial activities in the application, such as fund transfers, withdrawals, and evaluations of transaction features.

The topic modeling using the BERTopic algorithm on user reviews of the *Wondr by BNI* application showed good performance, achieving a coherence score of 0.5688, indicating a reasonably good topic quality. Additionally, the topic diversity score of 0.9667 reflects a very high level of topic diversity, with a large number of unique words used in the modeling process.

3.4. Aspect Labeling

Aspect labeling was performed using the cosine similarity method. The initial step involved converting review texts into numerical vectors using the Word2Vec model. Each review was then compared with a list of representative words for each aspect (such as *User Interface (UI)*, *authentication*, and *transaction*). If there was similarity between the words in the review and those in the aspect word list, the review was labeled according to the most relevant aspect. The representative words were obtained from the results of topic modeling and manual review analysis by the researchers.

Table 3. List of Aspects and Their Representations

Aspek	Representatif
UI (User Interface)	tampilan, simpel, menarik, keren, warna, colorful, desain, ui, antarmuka
Authentication	verifikasi, login, email, wajah, daftar, otp, kode, akun, logout
Transaction	transfer, tunai, tarik, saldo, transaksi, bayar, uang, rekening, topup, qris

The following table presents the results of aspect labeling using cosine similarity.

Table 4. Example of Aspect Labeling Results

Content_Stemmed	Aspect_Token	Aspect Label
[hidup, lebih, gampang, fiturnya, lengkap, desain, simpel, transaksi, cepat, banget, recommended, banget, sering, serba, praktis, sat, set, kayak]	{'Tampilan': ['desain', 'simpel'], 'Autentikasi': [], 'Transaksi': ['transaksi']}	UI, Transaction
[tolong, diperbaiki, masih, bug, daftar, masih, suruh, daftar, belum, bagus, tolong, diperbaiki, transfer, uang, tidak, bisa]	{'Tampilan': ['bagus'], 'Autentikasi': ['daftar', 'daftar'], 'Transaksi': ['transfer', 'uang']}	Authentication, Transaction
[seminggu, tidak, buka, giliran, buka, suruh, mulu, update, masa, bulan, kali, update]	{'Tampilan': [], 'Autentikasi': [], 'Transaksi': []}	NA

After this process, a total of 10,479 reviews remained that contained at least one of the three predefined aspects.

3.5. Sentiment Labeling

Sentiment labeling was carried out using the lexical window of context approach, which was implemented through three techniques: Lexicon InSet, RoBERTa1 = Indonesian RoBERTa Base Sentiment Classifier, and RoBERTa2 = RoBERTa Base IndoLEM Sentiment Classifier. The final sentiment label for each review was determined using the mode of these three outputs and then manually evaluated.

Table 5. Final Sentiment Labeling Results

Content	UI (User Interface)	Authentication	Transaction
Ngescan aja lama, gagal terus gk kebaca saat scan muka. Tolong diperbaiki masalah scan muka ngebug	NA	Negative	NA
Sangat simpel ,transaksi cuma pakai pin..data transaksi lebih jelas..sangat menarik dari segi tampilannya	Positive	NA	Positive
Daftarnya sangat mudah , transaksi juga lancar suka banget	NA	Positive	Positive
Tampilan doang bagus, transaksi kadang gagal tapi saldo kepotong hufft,	Positive	Na	Negative

3.6. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the initial stage of data analysis aimed at understanding the characteristics and structure of the dataset before proceeding to modeling or further analysis.

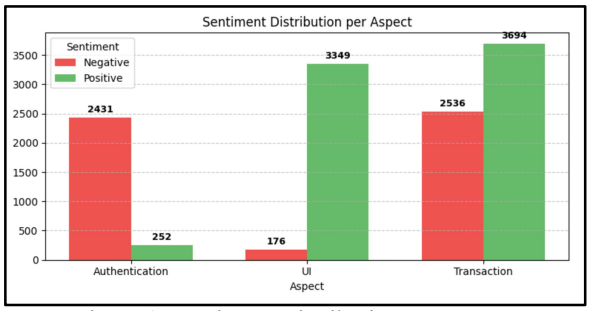

Figure 8. Sentiment Distribution per Aspect

This chart illustrates the sentiment distribution across the three main aspects of the *Wondr by BNI* application: authentication, UI (User Interface), and transaction. Authentication is heavily skewed toward negative sentiment, with 2,431 negative reviews and only 252 positive ones, indicating significant user frustration and recurring issues during the login or verification process. In contrast, the UI (User Interface)

stands out positively, receiving 3,349 positive reviews compared to only 176 negative ones, suggesting that users are highly satisfied with the app's visual design, layout, and overall usability. The transaction aspect shows the highest number of positive reviews (3,694) but also a considerable number of negative reviews (2,536), indicating that while many users are satisfied, there remain persistent issues that warrant attention. Overall, the data suggests that developers should prioritize improvements in the authentication and transaction features, while the UI aspect can be considered a strength of the application.

3.7. TF-IDF Weighting

To enable text data to be classified using machine learning, it is converted into numerical form through TF-IDF weighting. This process is automated using the TfidfVectorizer from the sklearn library.

3.8. Data Splitting

Data is split using the hold-out technique with an 80:20 ratio. After splitting, the data is further divided into two groups: the original data and the resampled data using SMOTE.

3.9. Model Design

In the model design, there are 10 testing scenarios consisting of 5 scenarios using the original data and 5 others using data balanced with the SMOTE method. These ten scenarios are designed to evaluate the model's performance in two separate classification stages: aspect classification and sentiment classification for each aspect.

Table 6. Model Testing Scenarios

No	Imbalance	Algoritman
1	Without Resampling	Support Vector Machine
2		Naive Bayes
3		Random Forest
4		Logistic Regression
5		*Stacking Ensemble*
6	SMOTE	Support Vector Machine
7		Naive Bayes
8		Random Forest
9		Logistic Regression
10		*Stacking Ensemble*

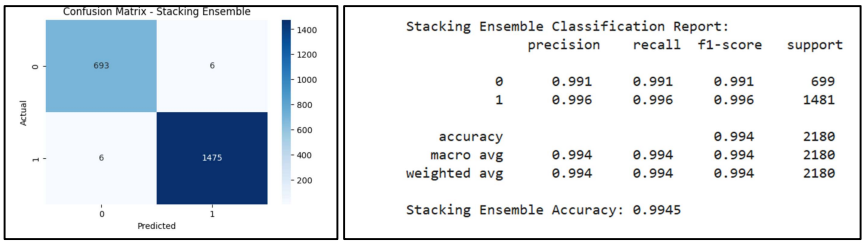

Figure 9. Example Confusion Matrix Stacking Ensemble

3.10. Model Evaluation

Model evaluation aims to measure a model's performance in assessing how well it completes. The following is an evaluation of the aspect model.

Table 7. Aspect Model Evaluation

No	Imbalance	Algoritma	Aspect	Accuracy	Precision	Recall	F1-Score
1	With	SVM	UI	0.984	0.988	0.976	0.982
			Authentication	0.988	0.992	0.984	0.973
			Transaction	0.987	0.984	0.989	0.986

2	out Resampling	Naive Bayes	UI	0.930	0.921	0.918	0.919
			Authentication	0.952	0.960	0.905	0.929
			Transaction	0.877	0.901	0.85	0.864
3		Random Forest	UI	0.981	0.982	0.974	0.978
			Authentication	0.983	0.988	0.964	0.975
			Transaction	0.986	0.986	0.985	0.985
4		Logistic Regression	UI	0.976	0.983	0.963	0.972
			Authentication	0.981	0.988	0.959	0.972
			Transaction	0.985	0.983	0.986	0.984
5		Stacking Ensemble	UI	0.994	0.994	0.994	0.994
			Authentication	0.995	0.993	0.993	0.993
			Transaction	0.991	0.989	0.991	0.990
6	SMOTE	SVM	UI	0.986	0.989	0.979	0.984
			Authentication	0.988	0.992	0.974	0.983
			Transaction	0.987	0.984	0.989	0.986
7		Naive Bayes	UI	0.849	0.830	0.872	0.879
			Authentication	0.929	0.887	0.93	0.906
			Transaction	0.928	0.930	0.919	0.924
8		Random Forest	UI	0.980	0.982	0.973	0.977
			Authentication	0.979	0.975	0.965	0.970
			Transaction	0.988	0.988	0.987	0.987
9		Logistic Regression	UI	0.983	0.987	0.975	0.980
			Authentication	0.989	0.992	0.977	0.984
			Transaction	0.981	0.978	0.984	0.981
10		Stacking Ensemble	UI	0.993	0.992	0.991	0.992
			Authentication	0.994	0.995	0.989	0.992
			Transaction	0.989	0.989	0.989	0.989

Based on the evaluation results, the Stacking Ensemble model delivered the best performance in classifying mobile banking service aspects. Due to the class imbalance in the data, the F1-score was used as the primary metric for evaluating model performance. Without applying any resampling method (such as SMOTE), the model achieved exceptionally high F1-scores: 99.4% for UI, 99.3% for Authentication, and 99% for Transaction. These results indicate that the model demonstrates excellent and consistent classification capabilities in identifying service aspects based on user reviews. The near-perfect performance highlights the effectiveness of Stacking Ensemble in combining the strengths of multiple base learners, enabling better pattern recognition compared to individual models.

Various sentiment testing scenarios were conducted on each aspect, and evaluation metrics were obtained, with F1-score being the primary metric due to significant class imbalance in sentiment for each aspect.

Table 8. Sentiment Model Evaluation

No	Imbalance	Algoritma	Sentiment	Accuracy	Precision	Recall	F1-Score
1	Without Resampling	SVM	UI	0.938	0.969	0.532	0.544
			Authentication	0.948	0.954	0.734	0.802
			Transaction	0.92	0.918	0.920	0.919
2		Naive Bayes	UI	0.912	0.956	0.539	0.550
			Authentication	0.935	0.967	0.511	0.504
			Transaction	0.908	0.905	0.909	0.907
3		Random Forest	UI	0.939	0.969	0.543	0.563
			Authentication	0.940	0.946	0.695	0.762
			Transaction	0.917	0.915	0.915	0.915
4		Logistic Regression	UI	0.936	0.968	0.521	0.524
			Authentication	0.942	0.970	0.696	0.766
			Transaction	0.917	0.914	0.917	0.917
5		Stacking Ensemble	UI	0.953	0.781	0.883	0.822
			Authentication	0.966	0.954	0.828	0.878
			Transaction	0.925	0.923	0.927	0.924
6	SMOTE	SVM	UI	0.945	0.89	0.605	0.655
			Authentication	0.952	0.957	0.754	0.82
			Transaction	0.918	0.916	0.917	0.917
7		Naive Bayes	UI	0.916	0.713	0.896	0.768
			Authentication	0.92	0.766	0.842	0.797
			Transaction	0.904	0.903	0.91	0.904
8		Random Forest	UI	0.94	0.766	0.701	0.728
			Authentication	0.95	0.905	0.779	0.828
			Transaction	0.905	0.903	0.906	0.904
9		Logistic Regression	UI	0.945	0.770	0.860	0.788
			Authentication	0.94	0.840	0.840	0.84
			Transaction	0.915	0.912	0.917	0.914
10		Stacking Ensemble	UI	0.945	0.972	0.585	0.631
			Authentication	0.957	0.958	0.783	0.846

Transaction	0.925	0.924	0.925	0.924

The Stacking Ensemble algorithm without resampling delivers the best performance in analyzing sentiment for each identified aspect of mobile banking services. The F1-score is considered the most relevant metric due to the significant class imbalance in sentiment data, especially in the UI (82.2%) and Authentication (87.8%) aspects. In the Transaction aspect, the model remains consistent with an F1-score of 92.4%, demonstrating its robustness in handling data variation.

In conclusion, Stacking Ensemble without resampling remains the best choice, offering optimal performance without requiring additional data manipulation. SMOTE may be more beneficial for other, more sensitive individual algorithms, but its impact is limited in Stacking Ensemble due to its strong base-model combination mechanism.

3.11. Model Interpretation

1. Interpretation of the UI Aspect Model

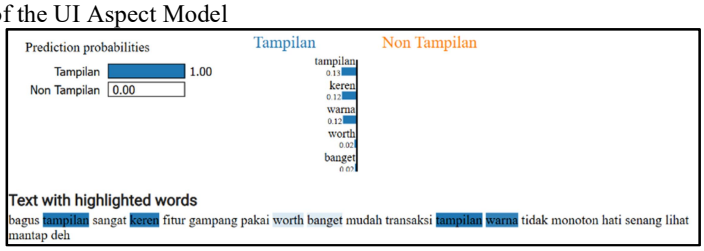

Figure 10. Interpretation of the UI Aspect Model

The model classifies the UI aspect with a probability of 1.00, identifying words such as "tampilan" (0.13), "keren" (0.12), and "warna" (0.12) as the main contributors. LIME helps explain the model's decision based on these relevant words.

2. Interpretation of the Authentication Aspect Model

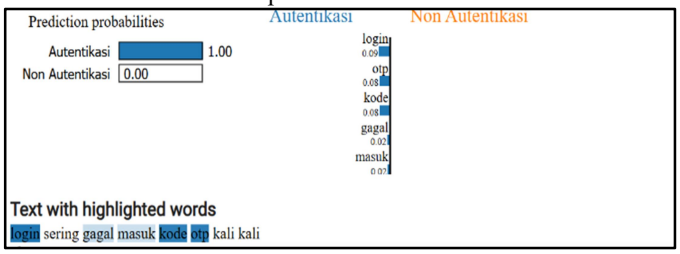

Figure 11. Interpretation of the Authentication Aspect Model

The model classifies the Authentication aspect with a probability of 1.00, identifying key words such as "login" (0.09), "otp" (0.08), and "kode" (0.08) as the main contributors. These words are highlighted in blue, indicating a strong influence on the model's decision.

3. Interpretation of the Transaction Aspect Model

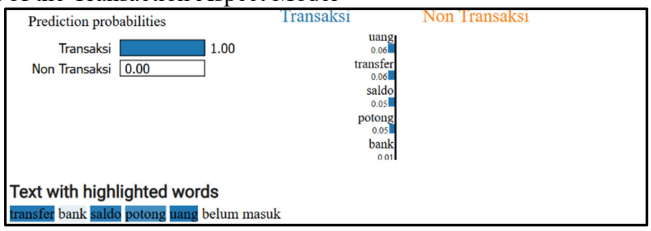

Figure 12. Interpretation of the Transaction Aspect Model

The model classifies the Transaction aspect with a probability of 1.00, with key words like "uang", "transfer", "saldo", and "potong" being the largest contributors. These words are highlighted in blue due to their strong impact on the model's decision.

4. Interpretation of the Positive Sentiment Model

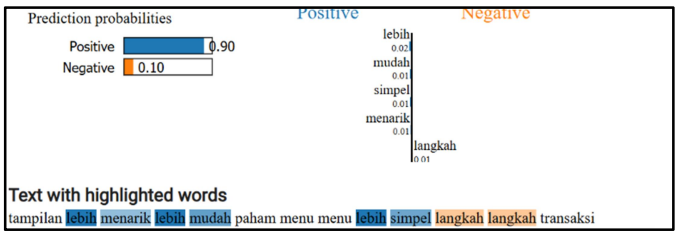

Figure 13. Interpretation of the Positive Sentiment Model

The model classifies the sentiment as Positive with a high probability (0.90), supported by words such as "lebih", "mudah", "simpel", and "menarik". The word "langkah" contributes slightly to the negative sentiment. Color highlights help indicate the influence of each word, and LIME explains how the model detects positive sentiment from words reflecting ease and attractiveness.

5. Interpretation of the Negative Sentiment Model

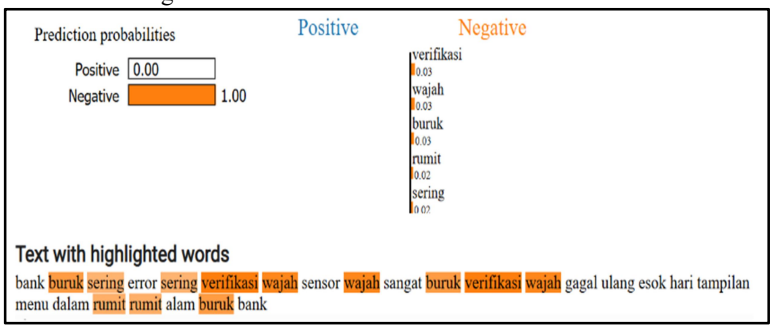

Figure 14.. Interpretation of the Negative Sentiment Model

The model classifies the sentiment as Negative with full probability (1.00), supported by words like "verifikasi", "wajah", "buruk", and "rumit", which have a strong impact. These words are highlighted in orange as they reflect negative experiences such as verification difficulties and complicated processes. LIME assists in explaining how the model recognizes negative sentiment based on these words.

3.12. Deployment

The best-performing model was exported in PKL format and integrated into an aspect-based sentiment classification website using the Flask framework for the deployment stage.

Figure 15. Website Interface – Analyze by Text

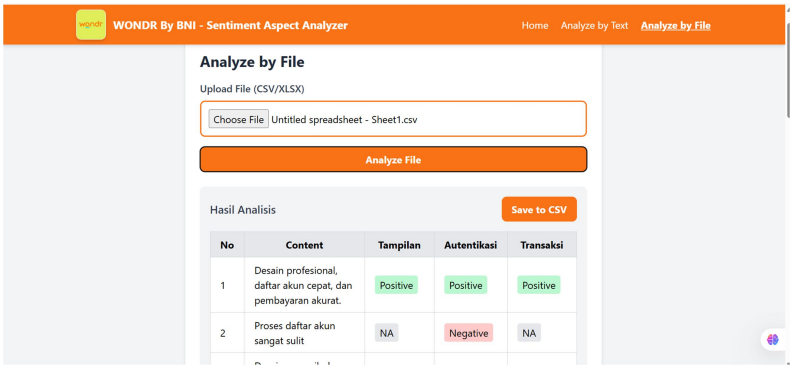

Figure 16. Website Interface – Analyze by File

These figures show that the system built using the best-performing model is capable of classifying the aspects and sentiments of user reviews effectively. This website simplifies the process for users to analyze multiple reviews at once without having to input data manually, making it useful for understanding user perceptions of mobile banking services more efficiently.

4. Conclusion

Based on the research findings, the Stacking Ensemble model without Resampling demonstrated the highest performance in aspect and sentiment classification compared to other individual models. The F1-Score for aspect classification reached 99.4% (UI), 99.3% (Authentication), and 99% (Transaction), while for sentiment classification, the F1-Scores were 82.2% (UI), 87.8% (Authentication), and 92.4% (Transaction). The application of SMOTE did not result in significant improvement, proving that the Stacking Ensemble model is already capable of handling data imbalance effectively. Furthermore, the use of LIME in review analysis showed that the model can provide transparent explanations and help interpret the contribution of words in aspect and sentiment classification, enhancing the model's interpretability and transparency. These findings underpin the development of a website capable of analyzing aspects and sentiments in app reviews, offering strategic insights for improving the Wondr by BNI application to better meet user needs.

References

[1] M. Khoirul, U. Hayati, and O. Nurdiawan, "Analisis Sentimen Aplikasi Brimo Pada Ulasan Pengguna Di Google Play Menggunakan Algoritma Naive Bayes," 2023.

[2] J. Iqbal and Isroq Urrahmah, "Pengaruh Kemudahan Dan Ketersediaan Fitur Terhadap Penggunaan Mobile Banking," 2021.

[3] Susi Setiawati, "Cashless Makin Digemari, Ini 5 Digital Banking Pilihan Warga RI," CNBC Indonesia, Jun.10,2024.https://www.cnbcindonesia.com/research/20240610063016-128-545113/cashless-makin-digemari-ini-5-digital-banking-pilihan-warga-ri (accessed Dec. 29, 2024).

[4] PT. Bank BNI, "Wondr by BNI." 2024. [Online]. Available: https://wondr.bni.co.id/. (Accessed: Dec. 29, 2024).

[5] Zefanya Aprilia, "Pengguna wondr Tembus 2 Juta, DPK BNI (BBNI) Bakal Meroket," CNBC Indonesia,Oct.11,2024.https://www.cnbcindonesia.com/market/20241011105508-17-578794/pengguna-uwondr-tembus-2-juta-dpk-bni--bbni--bakal-meroket (accessed Dec. 29, 2024).

[6] S. Fransiska, Rianto, and A. Irham Gufroni, "Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method," Sci. J. Informatics, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji

[7] A. Chandra Saputra, P. Raya Jln Hendrik Timang, P. Raya, P. Studi Teknik Informatika, F. Teknik, and U. Palangka Raya Jln Hendrik Timang, "Klasifikasi Rating Aplikasi Android Di Google Play Store Menggunakan Algoritma Gradient Boost Agus Sehatman Saragih, 2022 .

[8] N. B. Sidauruk, N. Riza, R. Nuraini, and S. Fatonah, "Penggunaan Metode SVM Dan Random Forest Untuk Analisis Sentimen Ulasan Pengguna Terhadap KAI Access Di Google Playstore," 2023.

[9] S. Amien, P. Perdana, T. Bharata Aji, and R. Ferdiana, "Aspect Category Classification dengan Pendekatan Machine Learning Menggunakan Dataset Bahasa Indonesia (Aspect Category Classification with Machine Learning Approach Using Indonesian Language Dataset)," 2021.

[10] A. Patel, P. Oza, and S. Agrawal, "Sentiment analysis of customer feedback and reviews for airline services using language representation model," *Procedia Computer Science*, vol. 218, pp. 2459–2467, 2023. [Online]. Available: https://doi.org/10.1016/J.PROCS.2023.01.221
[11] W. S. Dharmawan, "I N F O R M A T I K A Dalam Prediksi Penyakit Jantung," *Jurnal Informatika, Manajemen dan Komputer*, vol. 13, no. 2, 2021.
[12] Y. Pristyanto, "Penerapan Metode Ensemble untuk Meningkatkan Kinerja Algoritme Klasifikasi pada Imbalanced Dataset," 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/User+Knowledge
[13] A. K. Putri and H. Suparwito, "Uji Algoritma Stacking Ensemble Classifier pada Kemampuan Adaptasi Mahasiswa Baru dalam Pembelajaran Online," 2023.
[14] L. Maretva Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," 2022.
[15] S. George and D. Srividhya, "Aspect Based Sentiment Analysis of Restaurant Reviews using Ensemble Algorithms," 2021.
[16] R. Kalule, H. A. Abderrahmane, W. Alameri, and M. Sassi, "Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs," *Sci Rep*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-36096-2.
[17] Z. An, T. Xiong, Z. Zou, and H. Wan, "Aspect-based Sentiment Analysis with an Ensemble Learning Framework for Requirements Elicitation from App Reviews," *Journal of Internet Technology*, vol. 25, no. 7, pp. 1083–1090, Dec. 2024, doi: 10.70003/160792642024122507012.
[18] J. Khatib Sulaiman, N. Rizky Nuraeda, M. Liebenlito, and T. Edy Sutanto, "Explainable Sentiment Analysis pada Ulasan Aplikasi Shopee Menggunakan Local Interpretable Model-agnostic Explanations," *Indonesian Journal of Computer Science*.
[19] A. Munna, E. Zuliarso, U. Stikubank Jl Trilomba Juang No, and R. Artikel, "Interpretasi model Stacking Ensemble untuk analisis sentimen ulasan aplikasi pinjaman online menggunakan LIME," *AITI: Jurnal Teknologi Informasi*, vol. 21, no. 2, pp. 183–196, 2024.
[20] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," vol. 8, no. 1, pp. 147–156, 2021, doi: 10.25126/jtiik.202183944.
[21] Pavithra and Savitha, "Topic Modeling for Evolving Textual Data Using LDA, HDP, NMF, BERTOPIC, and DTM With a Focus on Research Papers," *Journal of Technology and Informatics (JoTI)*, vol. 5, no. 2, pp. 53–63, Apr. 2024, doi: 10.37802/joti.v5i2.618.
[22] Aishwarya Bhangale, "Introduction to Topic Modelling with LDA, NMF, Top2Vec and BERTopic," *Medium*, Mar. 08, 2023. https://medium.com/blend360/introduction-to-topic-modelling-with-lda-nmf-top2vec-and-bertopic-ffc3624d44e4 (accessed Jan. 07, 2025).
[23] A. Sanjaya, A. Bagus Setiawan, U. Mahdiyah, I. Nur Farida, A. Risky Prasetyo, and U. Nusantara PGRI Kediri, "Pengukuran Kemiripan Makna Menggunakan Cosine Similarity dan Basis Data Sinonim Kata Measurement of Meaning Similarity Using Cosine Similarity and Word Synonyms Database," vol. 10, no. 4, 2023, doi: 10.25126/jtiik.2023106864.
[24] A. Bandhakavi, N. Wiratunga, S. Massie, and R. Luhar, "Context extraction for aspect-based sentiment analytics: combining syntactic, lexical and sentiment knowledge." [Online]. Available: https://www.sentisum.com/, 2018.
[25] R. Nurhidayat and K. E. Dewi, "Komputa: Jurnal Ilmiah Komputer dan Informatika Penerapan Algoritma K-Nearest Neighbor dan Fitur Ekstraksi N-Gram dalam Analisis Sentimen Berbasis Aspek," vol. 12, no. 1, 2023.
[26] J. Khatib Sulaiman, N. Rizky Nuraeda, M. Liebenlito, and T. Edy Sutanto, "Explainable Sentiment Analysis pada Ulasan Aplikasi Shopee Menggunakan Local Interpretable Model-agnostic Explanations," *Indonesian Journal of Computer Science*.
[27] A. Munna, E. Zuliarso, U. Stikubank Jl Trilomba Juang No, and R. Artikel, "Interpretasi model Stacking Ensemble untuk analisis sentimen ulasan aplikasi pinjaman online menggunakan LIME," *AITI: Jurnal Teknologi Informasi*, vol. 21, no. 2, pp. 183–196, 2024.
[28] D. R. Sari, B. Matsaany, and M. Hamka, "Aspect Extraction of E-Commerce and Marketplace Applications Using Word2Vec and WordNet Path," Jurnal Teknik Informatika (Jutif), vol. 4, no. 4, pp. 787–796, Aug. 2023, doi: 10.52436/1.jutif.2023.4.4.726.