Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

Aspect Discovery and Sentiment Analysis of Bibit Application Reviews Using BERTopic and Transformer Models

Nyimas Princessa Syapeni[1], Apriansyah Putra[2*], Ari Wedhasmara[3], Ken Ditha Tania[4]
[1,2,3,4] Department of Information Systems, University of Sriwijaya, Indonesia
[1]09031382328170@student.unsri.ac.id, [2*]apriansyah@unsri.ac.id, [3]a_wedhasmara@unsri.ac.id, [4]kenya.tania@gmail.com

open access

***Corresponding Author**

Article History:
Submitted: 08-09-2026
Accepted: 16-09-2026
Published: 01-11-2026

Keywords:
Aspect Based Sentiment Analysis, BERTopic, Bibit, Class Weighted Loss, Transformer

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Digital investment applications generate large volumes of user reviews that contain valuable information about user experiences and perceptions. However, general sentiment analysis does not indicate the specific aspects that contribute to user's opinions. This study aims to identify user sentiment across aspects of the Bibit application using Aspect-Based Sentiment Analysis with a data-driven approach, compare the performance of Transformer models, and evaluate the effect of Class-Weighted Loss on sentiment classification. A total of 2.842 cleaned user reviews were analyzed using BERTopic to identify candidate aspects, which were subsequently interpreted and validated with the assistance of a Large Language Model (LLM). This process resulted in eight final aspects: UX/Usability, Transaction, Investment, Error/Performance, Account/Security, Feature/Product, Customer Service, and Promotion. Subsequently, 1,139 review-aspect instances were manually annotated as ground truth and used to evaluate five Transformer models, namely BERT, RoBERTa, IndoBERT, mBERT, and XLM-RoBERTa, under Unweighted and Class-Weighted Loss conditions. The results showed that IndoBERT with Unweighted Loss achieved the best performance, with an Accuracy of 85.09% and a Macro F1 of 82.74%. Class-Weighted Loss did not consistently improve performance across the evaluated models. At the aspect level, Transaction achieved a Macro F1 of 83.67%, while Customer Service obtained the lowest score of 60.00%. These findings indicate that the data-driven approach can identify relevant aspects from user reviews, while model selection and class-weighting strategies influence classification performance across aspects.

INTRODUCTION

Public participation in Indonesia's capital market has continued to increase, as reflected in the number of Single Investor Identifications (SIDs), which reached 30.27 million as of August 7, 2026, representing a 48.67% increase from the previous year (Otoritas Jasa Keuangan, 2026). This growth has been supported by digital investment applications such as Bibit, which simplify the investment process for users.

The widespread use of Bibit has generated a large number of user reviews on the Google Play Store, containing experiences, evaluations, complaints, and suggestions. These reviews provide valuable information for understanding user perceptions. However, their large volume makes manual analysis less efficient. Several studies have applied machine learning to analyze sentiment in Bibit reviews (Septiara & Fauzi, 2025; Yulianto & Andrianto, 2025). Nevertheless, these studies have primarily focused on general sentiment classification and have not identified the specific aspects of the application that contribute to users' sentiments.

Aspect-Based Sentiment Analysis (ABSA) provides a more detailed analysis by associating sentiment with specific aspects mentioned in reviews. Previous research (Serly Setyani, 2023) applied ABSA to Bibit using BERT/IndoBERT and three aspects, namely System, Service, and User Satisfaction, based on the DeLone-McLean model. However, these aspects were defined theoretically before the analysis, which may limit the identification of aspects that emerge directly from users' experiences. In contrast, studies on other applications have shown that a data-driven approach using BERTopic can identify topics based on patterns found in user reviews (Hamid et al., 2025).

To address this gap, this study applies a data-driven ABSA approach to Bibit application reviews. BERTopic is employed to identify candidate aspects, which are subsequently validated with the assistance of a Large Language Model (LLM). Sentiment associated with each aspect is then classified using five Transformer models: BERT, RoBERTa, IndoBERT, mBERT, and XLM-RoBERTa. The models are evaluated under both Unweighted and Class-Weighted Loss settings to examine their respective effects on sentiment classification performance.

From an Information Systems perspective, this approach provides not only an understanding of user's sentiments but also identifies the specific parts of the application associated with those sentiments. Such information can help reveal user's perceptions of system quality and their overall experience with the application. Therefore, this study aims

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

to identify user aspects and sentiments in the Bibit application using a data-driven approach, compare the performance of five Transformer models, and evaluate the effect of Class-Weighted Loss on sentiment classification.

LITERATURE REVIEW

Sentiment analysis is an approach used to identify and classify opinions in text into sentiment categories, such as positive, negative, or neutral. In application reviews, this approach helps provide a general understanding of user's perceptions of a service. Previous studies have applied machine learning methods to analyze user reviews of the Bibit application, demonstrating that sentiment classification can be used to identify overall patterns in user opinions (Yulianto & Andrianto, 2025). However, general sentiment analysis does not indicate which specific part of an application contributes to a particular sentiment. This limitation motivates the use of Aspect-Based Sentiment Analysis (ABSA), which associates sentiment with specific aspects within a review, allowing for more detailed and actionable analysis. In the Bibit application, ABSA has been applied using BERT/IndoBERT with three aspects referring to the System, Service, and User Satisfaction dimensions of the DeLone-McLean model (Serly Setyani, 2023).

The three aspects in (Serly Setyani, 2023) were predefined based on a theoretical framework before the reviews were analyzed, representing a theory-driven approach. This differs from a data-driven approach, in which aspects are identified from patterns that emerge directly from the data. Previous studies on other applications have demonstrated the potential of this approach. For example, topics such as login/authentication and error were identified directly from user review patterns using BERTopic in government service applications, without predefined categories (Hamid et al., 2025).

BERTopic is a topic modeling method that groups documents based on semantic similarity and generates keywords representing each topic, making it suitable for data-driven topic discovery in text collections (Egger & Yu, 2022). This approach has also been applied to digital identity applications (Dwiasnati et al., 2026), mobile banking applications (Hamid et al., 2025), and employment service applications (Putri, Pravitasari, & Winarni, 2026), where it successfully identified topics related to user-reported issues. However, topics generated by BERTopic require further interpretation to understand the underlying discussion themes before being used to define candidate aspects (Pinilih, Novita Dewi, & Alzami, 2026).

To support this interpretation process, Large Language Models (LLMs) can assist in understanding topic modeling results, as demonstrated by previous research using generative AI for topic interpretation (Yi, Oh, & Kim, 2025). In this study, the role of Google Gemini 3.6 Flash is limited to assisting in the validation of candidate aspects based on keywords and representative reviews, while the final decision remains with the researcher. Once the aspects are established, sentiment classification requires models capable of capturing contextual relationships in language. Transformer-based models are widely used for this purpose, including BERT, RoBERTa, IndoBERT, mBERT, and XLM-RoBERTa. Comparing these models provides a basis for determining their suitability for sentiment classification in Indonesian application reviews. In addition, class distribution may affect classification performance, as models tend to favor classes with more training instances. Class-Weighted Loss addresses this issue by assigning greater weights to underrepresented classes during training.

Overall, previous sentiment research on the Bibit application has focused on general sentiment analysis (Septiara & Fauzi, 2025; Yulianto & Andrianto, 2025) and ABSA with theoretically predefined aspects (Serly Setyani, 2023). Meanwhile, studies on other applications have shown that BERTopic can identify topics directly from patterns in user reviews (Dwiasnati et al., 2026; Hamid et al., 2025). These findings provide an opportunity to apply a data-driven approach to Bibit reviews. Accordingly, this study uses BERTopic to identify candidate aspects, an LLM to assist in their validation, and five Transformer models to classify sentiment for each aspect, while comparing Unweighted and Class-Weighted Loss to examine the effect of class weighting on model performance.

METHOD

Research Workflow

This study aims to identify the best-performing Transformer model for aspect-based sentiment analysis of user reviews of the Bibit application. The aspects are identified using a data-driven approach based on patterns emerging from the collected reviews. The research process consists of data collection, preprocessing, aspect identification using BERTopic, aspect validation with the assistance of a Large Language Model (LLM), and the establishment of eight final aspects. Manual sentiment annotation is then performed to construct the ground truth, which is used for training and evaluating five Transformer models. Each model is evaluated under two loss conditions, namely Unweighted Loss and Class-Weighted Loss. The experimental results are subsequently compared to determine the best-performing model, which is further evaluated for each final aspect. The overall research workflow is presented in Figure 1.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

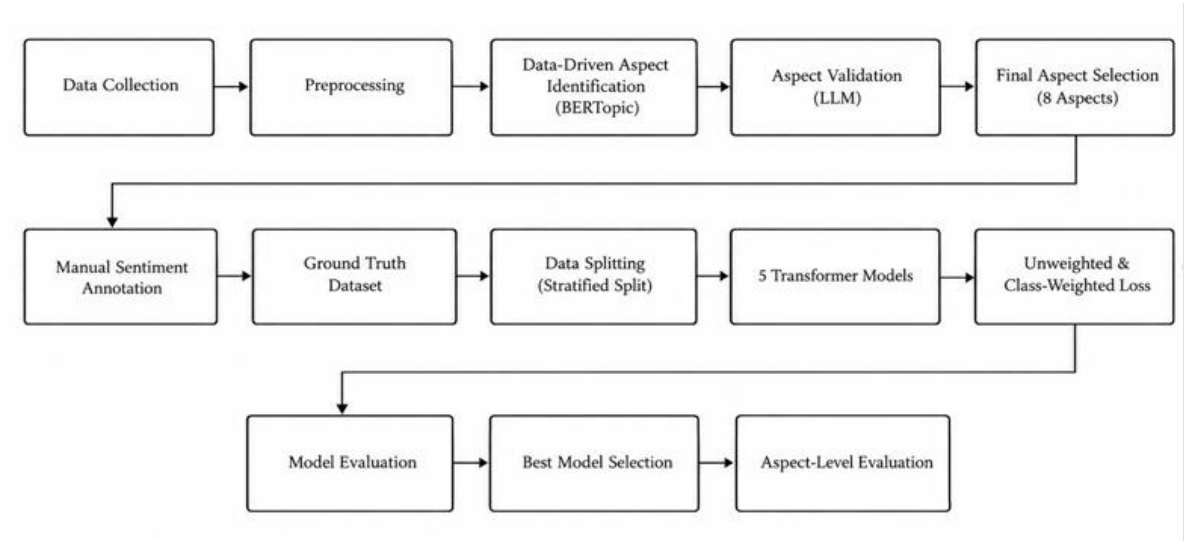


Figure 1. Research Workflow

Figure 1 illustrates the overall research workflow, including data collection, preprocessing, aspect discovery using BERTopic, aspect validation, sentiment annotation, and sentiment classification using Transformer models.

Research Data

The dataset consists of user reviews of the Bibit application collected from the Google Play Store for the 2024-2025 period. Data collection was conducted through a scraping process using Google Colab. The collected information included the review text, rating, and review date available on the application page. The resulting reviews were compiled into a single dataset and subsequently processed in the preprocessing stage.

Preprocessing

Preprocessing was performed to prepare the collected reviews for the subsequent aspect identification process. The preprocessing stage consisted of cleaning and case folding. The cleaning process removed elements considered irrelevant to text analysis, including URLs, mentions, hashtags, numbers, and punctuation marks. Case folding was then applied to convert all review text into lowercase to ensure consistency in text representation.

A quality check was also conducted to ensure that the dataset met the requirements for further analysis. From the initial 4,325 collected reviews, 52 empty reviews, 1.167 reviews containing fewer than two words or three characters, and 264 duplicate reviews were removed. After the filtering process, 2.842 clean reviews remained and were used as input for aspect identification using BERTopic.

Tokenization and stopword removal were not included as general preprocessing operations because these processes were specifically performed during text representation for BERTopic. Their implementation is described in the aspect identification stage.

Data-Driven Aspect Identification Using BERTopic

BERTopic is a topic modeling approach that uses Transformer-based embeddings to group documents according to their semantic similarity. This approach enables topics to be discovered directly from patterns within a text collection, making it suitable for a data-driven analysis (Egger & Yu, 2022).

In this study, the preprocessed reviews were first transformed into vector representations using a multilingual sentence embedding model. The resulting embeddings were then reduced in dimensionality using UMAP and grouped into clusters using HDBSCAN. Keywords representing each topic were subsequently extracted using class-based Term Frequency–Inverse Document Frequency (c-TF-IDF). The c-TF-IDF formulation used in this study is expressed as follows:

$$W_{(t,c)} = tf_{(t,c)} \times \log\left(1 + \frac{A}{tf_t}\right) \quad (1)$$

Information:
$W_{t,c}$= c-TF-IDF weight of term t in topic c.
$tf_{t,c}$= frequency of term t within topic c.
A= average number of words per topic.
tf_t= total frequency of term t across all topics.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

The calculation assigns a higher weight to words that are more representative of a particular topic. During the vectorizer stage, Indonesian tokenization and stopword removal were also applied to improve the relevance of the keywords generated for each topic.

For each topic, BERTopic produced a set of representative keywords along with several reviews that best reflected the topic, referred to as representative reviews. These outputs were examined to interpret the meaning of each topic (Uncovska, Freitag, Meister, & Fehring, 2023). Topics with similar meanings were then consolidated into candidate aspects. BERTopic was not used to assign sentiment labels; rather, its output served as the basis for identifying candidate aspects. The resulting candidates were subsequently validated before being established as the final aspects.

Aspect Validation and Finalization

The candidate topics generated by BERTopic were first interpreted based on their keywords and representative reviews. Since the resulting topics did not directly represent final aspects, further validation was required to assess their relevance and specificity to the research context. Previous research has also employed generative AI to assist in interpreting topic modeling results (Syukur Pangestu et al., 2025).

In this study, Google Gemini 3.6 Flash was used as a supporting tool to validate the candidate aspects obtained from BERTopic. The model received the topic keywords and representative reviews to assist in assessing the meaning and relevance of each candidate aspect. The validation was performed using a structured prompt with a temperature setting of 0.1, and the output was generated in JSON format. The generated responses were then reviewed by the researcher, who retained full responsibility for the final aspect selection. Gemini was not used for sentiment annotation or as a classification model. Following the interpretation, consolidation, and validation stages, eight final aspects were established for the subsequent sentiment annotation and classification.

Sentiment Annotation and Ground Truth

Following the establishment of the eight final aspects, manual sentiment annotation was performed on 1.139 review-aspect instances. Each instance was assigned to one of three sentiment categories: positive, negative, or neutral. Human annotation has also been applied in previous sentiment analysis studies, including research in which annotation guidelines were evaluated on a subset of the dataset before the complete annotation process (Pan & Meng, 2024).

The manually annotated data were subsequently treated as the ground truth for training and evaluating the five Transformer models.

Data Splitting

The 1,139 ground-truth instances were divided into training, validation, and test sets using a stratified split with a ratio of 80:10:10. This resulted in 911 instances for training, 114 for validation, and 114 for testing. The stratified strategy was applied to maintain the distribution of aspect-sentiment combinations across the three subsets.

The training set was used to fine-tune the models, while the validation set was used to monitor model performance during training and determine the best checkpoint. The test set was reserved for the final evaluation to measure model performance on data that had not been used during training.

Transformer Modeling

Five Transformer models were compared in this study, including BERT, RoBERTa, IndoBERT, mBERT, and XLM-RoBERTa. The models were selected to represent different characteristics, including general-purpose language models, a model developed specifically for Indonesian, and multilingual models. Previous research has also compared several Transformer models for Indonesian text classification (Riccosan & Saputra, 2025).

Each model was fine-tuned independently using the ground-truth dataset to classify sentiment into three categories: positive, negative, and neutral. The same experimental settings were applied across the models, while two different *loss* conditions were evaluated to obtain a consistent comparison of model performance.

Class-Weighted Loss

The distribution of sentiment classes in the ground-truth dataset was not completely balanced. Such an imbalance may cause a model to favor the majority class and reduce its ability to correctly recognize classes with fewer training instances (Cui, 2025).

To examine the effect of class weighting, each Transformer model was evaluated under two conditions: Unweighted Loss and Class-Weighted Loss. The combination of five Transformer models and two loss conditions resulted in 10 experimental scenarios. All scenarios were trained and evaluated under the same data partition to ensure a fair comparison.

For the Class-Weighted Loss condition, a weighted cross-entropy function was applied as follows:

$$L = -\sum_{c=1}^{C} w_c \, y_c \log(\hat{y}_c) \quad (2)$$

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

Information:
C= number of sentiment classes.
w_c= weight assigned to sentiment class c.
y_c= actual label for class c.
$\hat{y}_c$= predicted probability for class c.

The class weights were calculated based on the distribution of classes in the training set. Consequently, errors involving classes with fewer instances contributed more substantially to the loss value. This weighting strategy allows underrepresented classes to have a greater influence during model optimization. A similar class-weighting approach has been applied in IndoBERT-based sentiment classification to address class imbalance (Aprilah, Setiadi, & Herowati, 2025).

Experimental Scenarios and Evaluation

A total of 10 experimental scenarios were conducted, consisting of five Transformer models evaluated under two loss conditions. All scenarios used the same data partition to ensure that model performance could be compared under consistent conditions. The results were then compared to determine the model and loss configuration with the best overall performance.

Model performance was evaluated using Accuracy, Precision, Recall, F1-score, and Macro F1. The corresponding formulations are presented below.

- **Accuracy**

Accuracy represents the proportion of predictions that are correctly classified across the entire evaluation dataset. This metric provides an overall measure of the model's classification correctness.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

Information:
TP (True Positive): the number of instances correctly classified as belonging to the positive class.
TN (True Negative): the number of instances correctly classified as not belonging to the positive class.
FP (False Positive): the number of instances incorrectly classified as positive.
FN (False Negative): the number of positive instances incorrectly classified as negative.

- **Precision**

Precision measures the proportion of correct predictions within all instances predicted as positive. It indicates how reliable the model is when assigning an instance to the positive class.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Information:
TP (True Positive): positive instances that are correctly identified by the model.
FP (False Positive): instances predicted as positive even though they do not belong to the positive class.

- **Recall**

Recall measures the model's ability to identify instances that actually belong to the positive class. A higher recall indicates that fewer relevant instances are incorrectly missed.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Information:
TP (True Positive): positive instances correctly detected by the model.
FN (False Negative): positive instances that are incorrectly classified as negative.

- **F1-score**

F1-score combines Precision and Recall into a single measure by calculating their harmonic mean. This metric provides a balanced assessment when both the correctness of positive predictions and the ability to identify positive instances need to be considered.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (6)$$

Information:
Precision: the proportion of correct positive predictions among all positive predictions.
Recall: the proportion of actual positive instances successfully identified by the model.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

- **Macro F1**

Macro F1 measures the average F1-score across all sentiment classes. Each class contributes equally to the final score, regardless of the number of instances it contains.

$$MacroF1 = \frac{1}{C}\sum_{c=1}^{C} F\,1_c \quad (7)$$

Information:
C: the total number of sentiment classes.
$F1_c$: the F1-score obtained for the c-th sentiment class.

Macro F1 was selected as the primary evaluation metric because the sentiment classes in the dataset were imbalanced. By assigning equal importance to each class, this metric provides a more balanced representation of model performance across the sentiment categories (Aprilah et al., 2025).

After all 10 experimental scenarios were evaluated, the model and loss configuration with the highest overall performance were selected. The selected model was then evaluated separately for each of the eight final aspects. For this aspect-level evaluation, Macro F1 remained the primary metric to ensure that the performance of each sentiment class was considered equally.

RESULT

Initial Dataset Characteristics and Preprocessing

The research data were obtained from user reviews of the Bibit application. After reviewing and correcting the delimiter in the dataset file, 4,325 valid reviews were obtained, consisting of the content, score, and at attributes. The data then underwent preprocessing and a quality check to ensure that the dataset met the requirements for subsequent aspect modeling.

The quality check involved removing empty reviews, reviews with insufficient text length, and duplicate reviews. The results of the data cleaning process are presented in Table 1.

Table 1. Summary of Preprocessing and Data Cleaning

Processing Stage	Number of Reviews
Initial data after delimiter correction	4.325
Removed — empty/NaN reviews	52
Removed — excessively short reviews (<2 words / <3 characters)	1.167
Removed — duplicate reviews	264
Final cleaned data	**2.842**

The results indicate that 2.842 clean reviews remained from the initial 4,325 valid reviews after the quality checking process. These reviews were subsequently used as input for the data-driven aspect identification stage.

Data-Driven Aspect Identification Using BERTopic

A total of 2,842 cleaned reviews were analyzed to identify aspects using a data-driven approach. This approach was adopted so that the aspects could be derived from patterns present in the user reviews rather than being determined entirely from predefined assumptions.

The review texts were represented using the multilingual paraphrase-multilingual-MiniLM-L12-v2 sentence embedding model with a dimensionality of 384. The resulting representations were then processed using BERTopic through a combination of UMAP, HDBSCAN, and c-TF-IDF.

The process generated 35 topics. Among the 2.842 reviews, 646 reviews (22.73%) were classified as outliers, while 2.196 reviews (77.27%) were successfully assigned to topics.

Table 2. Topics Generated by BERTopic

Topic ID	Number of Reviews	Interpretation
0	311	Application Usability
1	247	Financial Transactions
2	236	Financial Transactions
3	205	General Satisfaction & Reputation
4	111	Application Usability
5	81	Application Usability
6	75	Account Registration/Verification
7	64	Account Registration/Verification
8	55	System Performance & Stability

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

9	52	Account Registration/Verification
10	45	Financial Transactions
11	42	Application Usability
12	42	System Performance & Stability
13	41	Investment Features & Products
14	38	Promotion & Referral Program

As shown in Table 2, the topics generated by BERTopic and their interpretations based on the characteristics of the reviews within each topic. The results show that the identified topics mainly relate to application usability, financial transactions, account registration and verification, system performance and stability, investment features and products, and promotion and referral programs. Several topics have similar interpretations, such as Topics 0, 4, 5, and 11, which are related to application usability, while Topics 1, 2, and 10 are related to financial transactions. These similarities indicate that multiple BERTopic topics can represent different parts of the same broader aspect. Therefore, the topics were further consolidated into more representative final aspect groups.

Consolidation of Topics into Aspect Groups

The topics generated by BERTopic were subsequently consolidated according to their semantic similarity. This process resulted in nine candidate aspect groups covering a total of 2.196 reviews that had been successfully clustered.

Table 3. Candidate Aspects Obtained from BERTopic Topic Consolidation

Aspect Group	Supporting Topics	Total Reviews
Usability & App Experience	9	683
Financial Transaction & Withdrawal	5	586
General Satisfaction & Trust	4	278
System Performance & Stability	7	261
Account Registration & Access	4	220
Investment Features & Products	2	64
Customer Service & Support	2	49
Promo & Referral Program	1	38
General Feedback	1	17

Table 3 presents the candidate aspect groups obtained by consolidating the topics generated by BERTopic based on their semantic similarity. The results show that Usability & App Experience had the largest number of reviews with 683 reviews, followed by Financial Transaction & Withdrawal with 586 reviews and General Satisfaction & Trust with 278 reviews. In contrast, General Feedback had the smallest number of reviews with 17 reviews. These results show how the generated topics were grouped into broader candidate aspects before the aspect validation process.

Validation of Aspect Groups Using an LLM

The consolidated aspect groups were subsequently validated using a Large Language Model (LLM), specifically Google Gemini 3.6 Flash. At this stage, Gemini served as an auxiliary tool for assessing the relevance and specificity of each candidate aspect within the fintech application domain. The LLM was not used as a sentiment annotator or classification model.

Table 4. LLM Validation Results for Candidate Aspects

Aspect Group	Validation Result	Decision
Usability & App Experience	Valid	Retained
Financial Transaction & Withdrawal	Valid	Retained
General Satisfaction & Trust	Invalid	Rejected
System Performance & Stability	Valid	Retained
Account Registration & Access	Valid	Retained
Investment Features & Products	Valid	Retained
Customer Service & Support	Valid	Retained
Promo & Referral Program	Valid	Retained
General Feedback	Invalid	Rejected

The validation results show that General Satisfaction & Trust and General Feedback were considered too broad because they did not refer to specific application features or functions. These two groups were therefore excluded, while the remaining groups were retained as the basis for defining the research aspects.

Based on data exploration, topic consolidation, and LLM-assisted validation, eight final aspects were established UX/Usability, Transaction, Investment, Error/Performance, Account/Security, Feature/Product, Customer Service, and Promotion. These aspects were consistently applied in the subsequent annotation and classification experiments.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

Manual Sentiment Annotation and Ground Truth Construction

After the final aspects had been established, manual sentiment annotation was performed by the researcher. Each review–aspect pair was assigned one of three sentiment labels: Negative, Neutral, or Positive. The resulting annotations were used as the ground truth for the Transformer-based classification and evaluation process.

The annotation process was conducted incrementally. Initially, 1.000 data instances were annotated, followed by additional annotation for aspect–sentiment combinations with relatively limited representation. This process resulted in a total of 1,139 ground-truth instances.

Table 5. Sentiment Distribution in the Ground Truth

Aspect	Negative	Neutral	Positive	Total
UX/Usability	132	73	110	315
Transaction	10	42	235	287
Investment	39	70	79	188
Error/Performance	84	21	7	112
Account/Security	0	35	44	79
Feature/Product	32	7	34	73
Customer Service	15	13	33	61
Promotion	0	12	12	24
Total	**312**	**273**	**554**	**1.139**

The sentiment distribution indicates that Positive was the largest class, consisting of 554 instances (48.6%), followed by Negative with 312 instances (27.4%) and Neutral with 273 instances (24.0%). Therefore, the dataset exhibited a moderate degree of class imbalance.

Dataset Splitting

The 1,139 ground-truth instances were divided using a stratified split based on the combination of aspect and sentiment, with a ratio of 80:10:10.

Table 6. Dataset Distribution

Dataset	Number of Instances	Purpose
Training	911	Training the model parameters
Validation	114	Monitoring performance and selecting the checkpoint
Test	114	Final model evaluation

The resulting dataset consisted of 911 training instances, 114 validation instances, and 114 test instances.

Training and Validation Results

Five Transformer models BERT, RoBERTa, IndoBERT, mBERT, and XLM-RoBERTa were evaluated under two loss conditions: Unweighted Loss and Class-Weighted Loss. Consequently, ten experimental scenarios were established and compared using the training, validation, and test sets. The best checkpoint for each scenario was selected based on the Macro F1 score obtained on the validation set.

Table 7. Macro F1 on the Validation Set per Epoch (%)

Model	Loss	Epoch 1	Epoch 2	Epoch 3
BERT	Unweighted	54.79	59.28	67.62
BERT	Weighted	58.45	63.31	65.11
RoBERTa	Unweighted	45.86	56.74	52.96
RoBERTa	Weighted	58.04	60.81	64.43
IndoBERT	Unweighted	84.02	83.91	79.33
IndoBERT	Weighted	81.18	83.67	80.01
mBERT	Unweighted	56.50	70.48	76.14
mBERT	Weighted	61.10	75.34	76.70
XLM-RoBERTa	Unweighted	54.93	63.17	64.14
XLM-RoBERTa	Weighted	52.97	71.83	67.58

As shown in Table 7**,** IndoBERT with Unweighted Loss achieved the highest validation Macro F1 in the first epoch, reaching 84.02%. The score subsequently decreased to 83.91% in the second epoch and 79.33% in the third epoch. Therefore, the first epoch was selected as the best checkpoint for the IndoBERT Unweighted scenario. The other models exhibited different Macro F1 patterns across epochs, indicating that training performance developed differently depending on the model and loss configuration.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

Comparison of the Ten Experiments on the Test Set

Following model training and best-checkpoint selection based on validation performance, all models were evaluated using the test data, which had not been used during training. The evaluation covered ten experimental scenarios combining five Transformer models with two loss configurations: Unweighted Loss and Class-Weighted Loss. Model performance was compared using Accuracy and Macro F1, with Macro F1 serving as the primary metric because it provides balanced consideration of each sentiment class.

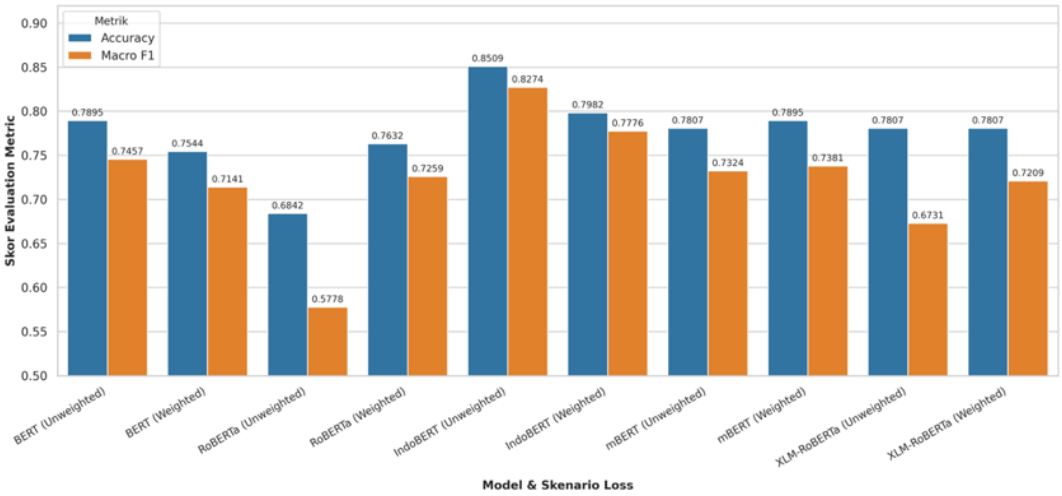


Figure 2. Comparison of Accuracy and Macro F1 Across All Models on the Test Set

The test results show in Figure 2 that IndoBERT with Unweighted Loss achieved the best overall performance, with an Accuracy of 85.09% and a Macro F1 of 82.74%. Its Macro F1 score was the highest among all ten experimental scenarios. These findings indicate that IndoBERT without class weighting provided the strongest overall sentiment classification performance on the research dataset.

Interestingly, the application of Class-Weighted Loss did not consistently improve model performance. An increase in Macro F1 was observed for RoBERTa, mBERT, and XLM-RoBERTa, whereas BERT and IndoBERT experienced performance reductions. Thus, the effect of Class-Weighted Loss appears to depend on the characteristics of each model rather than functioning as a universally beneficial strategy across all Transformer models.

Analysis of the Effect of Class-Weighted Loss

To examine the effect of Class-Weighted Loss on each model, the performance of the Unweighted Loss and Class-Weighted Loss scenarios was compared. The comparison focused on Macro F1 to assess changes in model performance after class weighting was introduced. The results are presented in Figure 3.

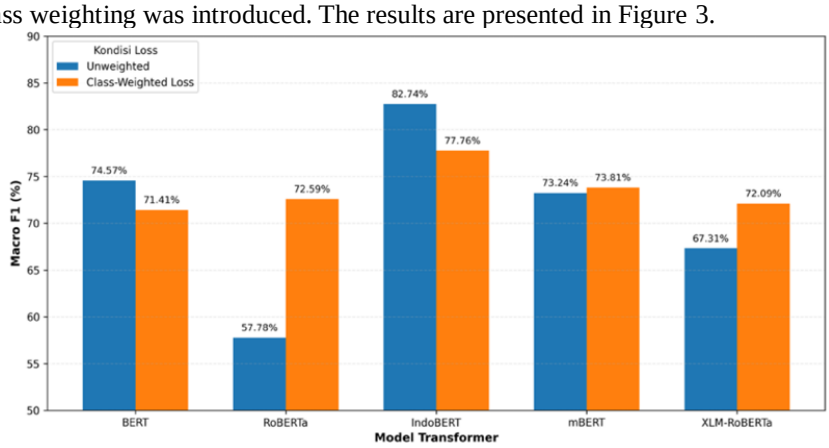


Figure 3. Comparison of Macro F1 Before and After Applying Class-Weighted Loss

Figure 3 demonstrates that Class-Weighted Loss did not produce a consistent improvement across all models. Its effectiveness varied according to the model being used. The observed pattern suggests that class weighting tended to benefit models that initially had greater difficulty recognizing the neutral class, particularly RoBERTa and XLM-RoBERTa. Under the Unweighted condition, these models recorded low neutral recall values of 14.81% and 18.52%, respectively. In contrast, models that already exhibited relatively balanced performance across sentiment classes, such

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

as BERT and IndoBERT, experienced less favorable results after class weighting was applied. Therefore, Class-Weighted Loss should be considered according to the characteristics of the model and dataset rather than applied uniformly to every Transformer model.

Analysis of the Best Model: IndoBERT Unweighted

Based on the evaluation of the ten experimental scenarios, IndoBERT with Unweighted Loss achieved the highest Macro F1 score of 82.74% and an Accuracy of 85.09%, making it the best-performing model configuration. To further examine its ability to classify each sentiment class and assess the effect of Class-Weighted Loss, a confusion matrix comparison was performed between the two IndoBERT configurations.

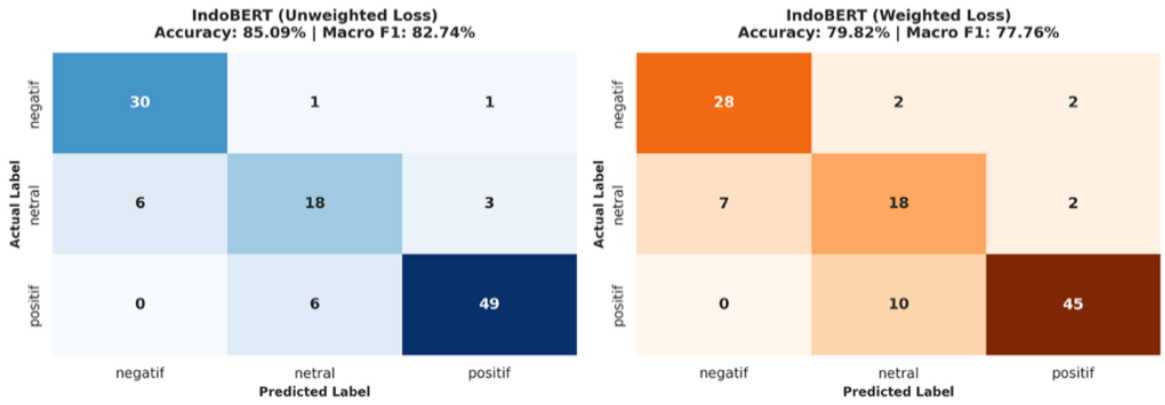


Figure 4. Confusion Matrix Comparison of IndoBERT with Unweighted and Weighted Loss

As shown in Figure 4, IndoBERT with Unweighted Loss correctly classified most of the negative and positive instances, with 30 out of 32 negative instances and 49 out of 55 positive instances correctly identified. For the neutral class, 18 out of 27 instances were correctly classified. Most classification errors occurred when neutral instances were predicted as either negative or positive.

Under the Class-Weighted Loss condition, the number of correctly classified negative instances decreased to 28, while correctly classified positive instances decreased to 45. The number of correctly classified neutral instances remained at 18. This result corresponds with the decrease in Macro F1 from 82.74% to 77.76%, indicating that Class-Weighted Loss did not improve IndoBERT's performance compared with the Unweighted Loss configuration.

Evaluation of the Best Model by Aspect

After IndoBERT with Unweighted Loss was selected as the best-performing configuration, a further evaluation was conducted for each of the eight aspects to examine variations in model performance across aspects. Macro F1 was retained as the primary evaluation metric because it considers the performance of each sentiment class equally. The results are presented in Figure 5.

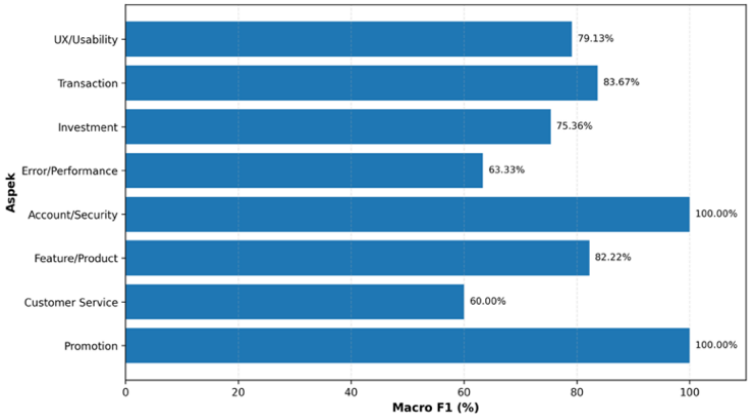


Figure 5. Macro F1 Performance of IndoBERT Unweighted Across Individual Aspects

The results show that the model performance varied across aspects. Transaction obtained the highest performance, while Customer Service recorded the lowest Macro F1 at 60.00%. Account/Security and Promotion achieved a Macro F1 of 100%. For Promotion, this result was influenced by the ground-truth distribution, which contained no negative instances. Therefore, the perfect score does not necessarily indicate better classification performance and should be interpreted cautiously.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

DISCUSSION

The data-driven aspect discovery process identified eight aspects commonly discussed in Bibit user reviews, namely UX/Usability, Transaction, Investment, Error/Performance, Account/Security, Feature/Product, Customer Service, and Promotion. The aspects were derived from patterns found in the reviews and then validated before being included in the sentiment analysis. This process ensured that the selected aspects were based on issues actually raised by users.

IndoBERT with Unweighted Loss achieved the highest overall performance, with an accuracy of 85.09% and a Macro F1 score of 82.74%. The results from Class-Weighted Loss varied across the five models. RoBERTa, mBERT, and XLM-RoBERTa showed higher Macro F1 scores after weighting, while BERT and IndoBERT showed lower scores. Under the Unweighted condition, RoBERTa and XLM-RoBERTa had low recall for the neutral class, at 14.81% and 18.52%, respectively. Their neutral-class recall improved after class weighting. However, IndoBERT's Macro F1 decreased from 82.74% to 77.76%. This result is in line with (Cui, 2025), who found that weighted loss did not produce consistent improvements across Transformer models. In that study, weighted loss improved BERT and BART, while RoBERTa showed only a slight change and a small decrease in Micro F1. These findings suggest that Class-Weighted Loss does not always provide better performance or fully address minority-class problems. Similar results were also reported by Ganganwar and Rajalakshmi (Ganganwar & Rajalakshmi, 2024), who showed that class-weighted cross-entropy did not consistently address the class imbalance problem in aspect-based sentiment classification.Performance also differed across aspects when using IndoBERT.

Transaction obtained a Macro F1 score of 83.67%, whereas Customer Service had the lowest score at 60.00%. This difference may be related to the amount of annotated data available for each aspect. Account/Security and Promotion both achieved a Macro F1 score of 100%, but these results should be considered carefully because the number of test instances for both aspects was relatively small.

CONCLUSION

This study successfully conducted aspect-based sentiment analysis of Bibit application reviews using a *data-driven* approach. BERTopic and LLM-assisted validation identified eight final aspects from 2.842 cleaned reviews, while 1.139 review-aspect instances were manually annotated and used to evaluate five Transformer models across ten experimental scenarios. IndoBERT with Unweighted Loss achieved the best overall performance, with 85.09% Accuracy \and 82.74% Macro F1. Class-Weighted Loss did not consistently improve performance across models. At the aspect level, Transaction achieved the highest Macro F1 of 83.67%, while Customer Service obtained the lowest score of 60.00%.

These results show that data-driven aspect discovery combined with Transformer-based classification can provide aspect-level insights from user reviews. Future research should expand annotated data, particularly for aspects with limited instances, and explore other class imbalance strategies and Transformer models to improve classification performance across different aspects.

REFERENCES

Aprilah, T., Setiadi, D. R. I. M., & Herowati, W. (2025). Enhancing Aspect-Based Sentiment Analysis via Hugging Face Fine-Tuned IndoBERT. *Journal of Applied Informatics and Computing*, *9*(6), 3821–3830. https://doi.org/10.30871/jaic.v9i6.11409

Cui, X. (2025). *xiacui at SemEval-2025 Task 11: Addressing Data Imbalance in Transformer-Based Multi-Label Emotion Detection with Weighted Loss*. Retrieved from https://aclanthology.org/2025.semeval-1.37/

Dwiasnati, S., Arya Putri, C. D., Sandiwarno, S., Yuliarty, P., Gunawan, W., Devianto, Y., & Avrizal, R. (2026). BERTopic Driven Topic Discovery for User Experience Evaluation: A Sentiment Study on Google Play Reviews of a Digital Identity Application. *2026 International Conference on Enterprise and Industrial System (ICOEINS)*, 1–6. IEEE. https://doi.org/10.1109/ICOEINS69996.2026.11605880

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Ganganwar, V., & Rajalakshmi, R. (2024). Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification. *Journal of Information and Telecommunication*, *8*(2), 167–188. https://doi.org/10.1080/24751839.2023.2270824

Hamid, R. B., Andriansyah, C., Sensuse, D. I., Lusa, S., Elisabeth, D., & Safitri, N. (2025). Sentiment Analysis and Topic Modeling for Discovering Knowledge in Indonesian Mobile Government Applications. *Jurnal Teknik Informatika (Jutif)*, *6*(5), 3188–3203. https://doi.org/10.52436/1.jutif.2025.6.6.4991

Otoritas Jasa Keuangan. (2026, August 10). Siaran Pers: Penguatan Integritas Pasar Modal Jadi Prioritas Utama, 49 Tahun Diaktifkannya Kembali Pasar Modal Indonesia, OJK Luncurkan ETF Emas. *Otoritas Jasa Keuangan (OJK)*.

Pan, Q., & Meng, Z. (2024). Hybrid Uncertainty Calibration for Multimodal Sentiment Analysis. *Electronics*, *13*(3),

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9635

662. https://doi.org/10.3390/electronics13030662

Pinilih, N. W., Novita Dewi, I., & Alzami, F. (2026). An Integrated Topic–Sentiment Analysis of User Reviews in Vidio Application Using BERTopic and IndoRoBERTa. *Journal of Applied Informatics and Computing*, *10*(3), 2685–2698. https://doi.org/10.30871/jaic.v10i3.12952

Putri, N. A., Pravitasari, A. A., & Winarni, S. (2026). Topic-based sentiment analysis of Jamsostek mobile application reviews using the BERTopic method. *Commun. Math. Biol. Neurosci.*, *2026*(0). https://doi.org/10.28919/cmbn/9682

Riccosan, R., & Saputra, K. E. (2025). Multilabel classification sentiment analysis on Indonesian mobile app reviews. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *14*(5), 4226. https://doi.org/10.11591/ijai.v14.i5.pp4226-4234

Septiara, F. P., & Fauzi, R. A. (2025). Analisis Sentimen Pelanggan Terhadap Aplikasi Investasi di Indonesia. *Jurnal Inovasi Global*, *3*(7), 1136–1151. https://doi.org/10.58344/jig.v3i7.382

Serly Setyani. (2023). Multi Aspect Sentiment Analysis of Mutual Funds Investment App Bibit Using BERT Method. *International Journal on Information and Communication Technology (IJoICT)*, *9*(1), 44–56. https://doi.org/10.21108/ijoict.v9i1.718

Syukur Pangestu, A., Christian, E., Lestari, A., Nyaho, T., Yos Sudarso, J., & Raya, P. (2025). Aspect-Based Sentiment Analysis pada Ulasan Pengguna Aplikasi Mobile JKN Menggunakan Model Berbasis Transformer. *JOINTECOMS (Journal of Information Technology and Computer Science) p-ISSN: 2798-284X*, *5*(4), 2798–3862. https://doi.org/10.47111/jointecoms.v5i4

Uncovska, M., Freitag, B., Meister, S., & Fehring, L. (2023). Rating analysis and BERTopic modeling of consumer versus regulated mHealth app reviews in Germany. *Npj Digital Medicine*, *6*(1), 115. https://doi.org/10.1038/s41746-023-00862-3

Yi, J., Oh, Y. K., & Kim, J.-M. (2025). Unveiling the drivers of satisfaction in mobile trading: Contextual mining of retail investor experience through BERTopic and generative AI. *Journal of Retailing and Consumer Services*, *82*, 104066. https://doi.org/10.1016/j.jretconser.2024.104066

Yulianto, M. A., & Andrianto, R. (2025). Analisa Kinerja Algoritma Supervised Learning pada Sentimen Ulasan Aplikasi Investasi Online Bibit. *RIGGS: Journal of Artificial Intelligence and Digital Business*, *4*(3), 1486–1496. https://doi.org/10.31004/riggs.v4i3.2168

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.