Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

Fine-Tuning the Gemini 1.5 Flash Large Language Model for User Perception Classification in BSI Mobile Application Reviews

Rio Fidelis[1], Vicraj[2], Dea Monica Bangun[3], Nur Mayanti[4], Evta Indra[5]
Universitas Prima Indonesia

Article Info

Keywords:
Large Language Model,
Fine-Tuning,
Gemini 1.5 Flash,
Perception Classification,
Sentiment Analysis,
IndoBERT,
Google Cloud Vertex AI.

ABSTRACT

he growing volume of user reviews on digital platforms such as the Google Play Store presents a major challenge in automatically understanding user perceptions, especially due to the unstructured, varied, and highly subjective nature of the text data. Manual analysis at this scale is inefficient and prone to bias. To address this issue, this study applies fine-tuning on the Large Language Model (LLM) Gemini 1.5 Flash to automatically classify user perceptions of the BSI Mobile application. Perceptions are categorized into three classes: Very Poor, Fair, and Excellent. A total of 120,000 reviews were collected via web scraping and processed through cleaning, normalization, automatic labeling using the IndoBERT model, and conversion into JSONL format for fine-tuning on the Google Cloud Vertex AI platform. Evaluation results show an accuracy of 63.41% for perception classification and 67.31% for sentiment classification, with F1-scores of 28.82% and 28.75%, respectively. The model demonstrated better accuracy in identifying positive perceptions, while neutral or ambiguous reviews remained a challenge. Consistency analysis between predicted perceptions and user ratings showed a match rate of 83.81%. This study demonstrates that the fine-tuned Gemini 1.5 Flash is an effective solution for text-based perception classification and holds strong potential for broader application in user opinion analytics systems.

Corresponding Author:
Evta Indra
Universitas Prima Indonesia
evtaindra@unprimdn.ac.id

INTRODUCTION

The rapid advancement of *Artificial Intelligence* (AI), particularly in *Natural Language Processing* (NLP), has significantly transformed various sectors such as digital banking, customer experience assessment, and public sentiment analysis (Beurer-Kellner, Fischer and Vechev, 2023; Liu *et al.*, 2023). In this context, the emergence of *Large Language Models* (LLMs) like GPT-4 and Gemini 1.5 Flash has redefined how systems interpret and classify unstructured text with remarkable accuracy and contextual depth (Gerlich, Elsayed and Sokolovskiy, 2023; Olujimi and Ade-Ibijola, 2023). A persistent challenge in NLP involves accurately categorizing user impressions, which often feature informal language, sarcasm, emotional tone, and complex linguistic variations (Fatouros *et al.*, 2024). In the realm of digital banking, understanding user perception is crucial for enhancing service quality and maintaining trust (Sottana *et al.*, 2023; Malik and Bilal, 2024). BSI Mobile, developed by Bank Syariah Indonesia (BSI), receives a large number of user reviews each month through the

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

Google Play Store, reflecting diverse experiences related to interface usability, transaction performance, system reliability, and overall satisfaction. However, due to the high volume and complexity of this data, manual classification is inefficient and vulnerable to subjective interpretation. Consequently, an automated, scalable, and intelligent solution is essential for consistently and accurately identifying user perception (Al-Baity *et al.*, 2022; Arora and Banerji, 2024).

Previous studies have investigated sentiment and perception classification using *Machine Learning* (ML) and *Deep Learning* (DL) methods, including *Support Vector Machines* (SVM), *Naive Bayes*, and *Long Short-Term Memory* (LSTM) (Arifiyanti, Shantika and Syafira, 2023; Trani and Tran, 2024). Although these techniques offer foundational solutions, they often fall short in capturing the semantic depth of Indonesian text, especially when training data is limited. In addition, conventional models generally lack the flexibility and language-specific adaptation needed to handle the nuances of user-generated content in Bahasa Indonesia (Sodik, Nur Zaida and Zulmiati, 2022). To address these limitations, this study applies fine-tuning on Gemini 1.5 Flash, an advanced *Large Language Model* (LLM) developed by Google, to classify user perceptions in reviews of the BSI Mobile application. The research uses a dataset of 120,000 user reviews collected through web scraping from the Google Play Store, which are then preprocessed, normalized, and labeled using the IndoBERT model (Ahmad *et al.*, 2022). The fine-tuning process is conducted on this domain-specific dataset using *Google Cloud Vertex AI* to enhance classification performance for texts written in Indonesian Language.

The novelty of this research lies in the integration of three key components: (1) the application of Gemini 1.5 Flash with fine-tuning on user reviews from an Indonesian banking application, (2) the structured categorization of perceptions into three sentiment based classes, namely Very Bad, Fair, and Very Good, and (3) the use of cloud based model optimization through *Google Cloud Vertex AI*. These contributions are expected to provide fresh insights into the utilization of large language models for localized, perception-oriented classification tasks and to support financial institutions in gaining a more comprehensive understanding of customer experiences (Wong *et al.*, 2023; Fahrani and Aryanto, 2024).

Therefore, this study aims to (1) implement fine-tuning of the Gemini 1.5 Flash model using a labeled dataset of BSI Mobile reviews and (2) evaluate the model's Performance in accurately classifying user perceptions using standard metrics, including accuracy, precision, recall, and F1-score.

METHODS

This research adopts a quantitative experimental approach to evaluate the effectiveness of fine-tuning the Gemini 1.5 Flash model in classifying user perceptions from BSI Mobile application reviews. The process was carried out systematically through several stages, described below, to ensure clarity, reproducibility, and objectivity

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

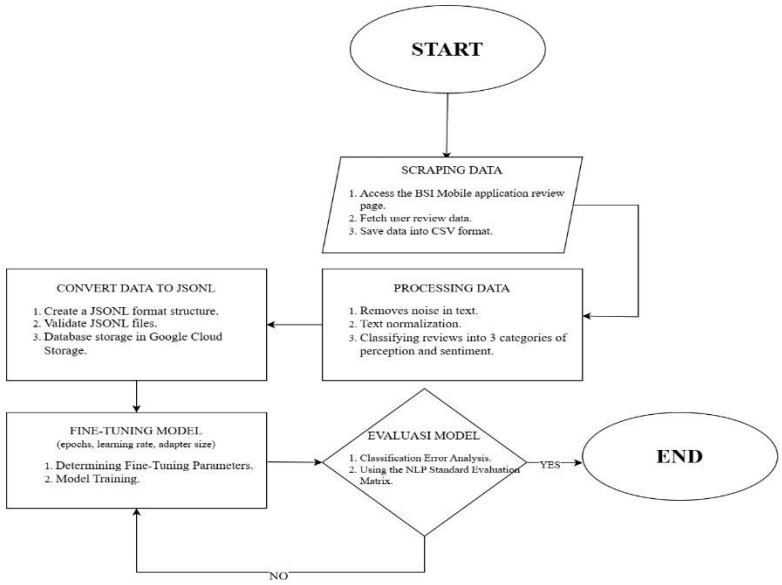


Figure 1. Flowchart Diagram

Data Collection

The dataset consisted of 120,000 user reviews collected from the Google Play Store using the Google-play-scraper library in Python. The reviews were retrieved specifically for the BSI Mobile application to ensure domain relevance. Data was collected in raw textual form, containing a mix of formal and informal Indonesian language, emojis, and unstructured expressions.

Tabel 1. BSI Mobile Data Scraping Sample

Content	Score	at
Sudah 7 hari ajukan pengaduan pelayanan kurang memuaskar saldo terpotong tapi tidak sukses,	1	2025-03-03 07:59:50
Aplikasi yang sangat bagus dan keren	5	2025-03-03 06:03:05
bank elit sistem sulit atm tiba tiba di blokir	1	2025-03-03 04:25:10
Apanya demi kenyamanan ketika cek saldo disuruh ke aplikasi lair dan DIPAKSA	1	2025-03-03 04:14:12
Pertama ok saja makin kesini makin ribet,tetiba akunnya tidak bisa di akses minta download app yang baru,di download ngisi data udah sesuai malah tidak valid terus.	2	2025-03-02 18:44:28

Data Preprocessing

Preprocessing was conducted to remove noise, normalize language use, and eliminate duplicates to prepare the text for model training. Specifically, non-linguistic elements such as emojis, HTML tags, special characters, and extra white spaces were cleaned from the text.

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

Informal and non-standard words were normalized into proper Indonesian. Identical and near-identical entries were removed to ensure data uniqueness.

Following the cleaning and normalization process, the *IndoBERT* pretrained model (*cahya/bert-base-indonesian-1.5G-sentiment-classification*) was utilized to assign perception labels to each review. Unlike standard sentiment analysis, this study classified reviews into perception-based categories to reflect the overall user experience. The reviews were automatically labelled into three categories: Very Good (*Sangat Baik*), Fair (*Cukup*), and Very Bad (*Sangat Buruk*). These labels aimed to capture user expectations and satisfaction more accurately within the context of a digital banking application.

Tabel 2. Sample Data Preprocessing

cleaned_review	score	at	sentiment	perception
sudah 7 hari ajukan pengaduan pelayanan kurang memuaskan saldo terpotong tapi tidak sukses	1	2025-03-03 07:59:50	negative	Sangat Buruk
aplikasi yang sangat bagus dan keren	5	2025-03-03 06:03:05	positive	Sangat Baik
bank elit sistem sulit atm tiba di blokir	1	2025-03-03 04:25:10	negative	SangatBuruk
apanya demi kenyamanan ketika cek saldo disuruh ke aplikasi lain dan dipaksa	1	2025-03-03 04:14:12	neutral	Cukup
pertama oke saja semakin ke sini semakin sulit, tetiba akunya tidak bisa di akses minta unduh apa yang baru, di unduh mengisi data sudah sesuai malah tidak valid terus	2	2025-03-02 18:44:28	negative	Sangat Buruk

Dataset Formatting

The dataset used in this study consisted of user reviews that had previously undergone cleaning and labelling. For training and evaluation, the dataset was split into three subsets: 70% was allocated for training, 15% for validation, and the remaining 15% for testing. This split ensured that the model could be both trained effectively and evaluated fairly (Catania *et al.*, 2022).

After the split, the dataset was converted into JSONL format, compatible with *Google Cloud Vertex AI* for fine-tuning large language models. The JSONL files were then uploaded to *Google Cloud Storage*, where they served as the primary data source during the fine-tuning process on *Vertex AI*.

However, due to platform constraints on *Vertex AI*, the number of validation examples used during the fine-tuning process was limited to a maximum of 5,000. Consequently, only 5,000 samples from the 15% validation subset were utilized, while the remainder was

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

excluded from the fine-tuning phase. This adjustment was necessary to meet platform requirements while maintaining the intended proportions for training and testing data.

The fine-tuning process was conducted using *gemini-1.5-flash-002* through the *Google Cloud Vertex AI* platform. The training was deployed in the *Asia-southeast1* region to ensure optimal latency and Performance. The configuration applied during fine-tuning included a learning rate multiplier of 1.0, an adapter size of 16, and a total of 15 training epochs. These hyperparameter settings were chosen to balance training efficiency and model accuracy.

Evaluation Metrics

The classification model was evaluated using four main metrics: accuracy, precision, recall, and F1-score. These metrics assessed how effectively the fine-tuned Gemini 1.5 Flash model classified user perceptions and sentiments based on review data from the BSI Mobile application.

The formula used to calculate accuracy is shown in equation (1), where the value is derived by dividing the sum of true positives (TP) and true negatives (TN) by the total number of instances, including false positives (FP) and false negatives (FN):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Equation (2) is used to determine the precision. Precision represents the proportion of correctly predicted positive observations out of all predicted positives, calculated as follows:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall, as described in equation (3), reflects the model's ability to retrieve all relevant positive instances. It is obtained by dividing the true positive count by the total of true positives and false negatives:

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

The F1-score, or F-measure, provides a single metric that balances both precision and recall. It is computed as the harmonic mean of the two values using equation (4):

$$F1 - Score = 2x \frac{Precision\ X\ Recall}{Precision + Recall} \quad (4)$$

The evaluation of the model's overall Performance took into account each of these metrics to ensure a comprehensive understanding of its effectiveness. The classification system is considered optimal when these values are close to 1, indicating a high level of prediction accuracy across all classes (Kaur and Sandhu, 2023).

RESULT AND DISCUSSION

The fine-tuning process of the Gemini 1.5 Flash model using a dataset of 120,000 user reviews from the BSI Mobile application yielded promising outcomes across several evaluation stages. These results are systematically presented below to illustrate the model's capability in classifying user perception and sentiment. Additional tuning was performed on the dataset that had been converted into JSONL format before uploading it to the model. Before the fine-tuning process was completed, the dataset was divided into 83,607 training data, 5,000 validation data (a subset of the full 15% due to limitations in Vertex AI), and 17,916 testing data. This division follows standard practice in machine learning, where

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

generally, the dataset is allocated 70% for training, 15% for validation, and 15% for testing. Training is used to train the model, validation helps in hyperparameter optimization and prevents overfitting, while testing serves to evaluate the final performance of the model against new data. After the fine-tuning process is complete, the data distribution is analyzed based on the number of input tokens, output tokens, and the number of messages per instance to ensure data balance and improve model accuracy.

The following are the results obtained from the data distribution of the number of input tokens per instance:

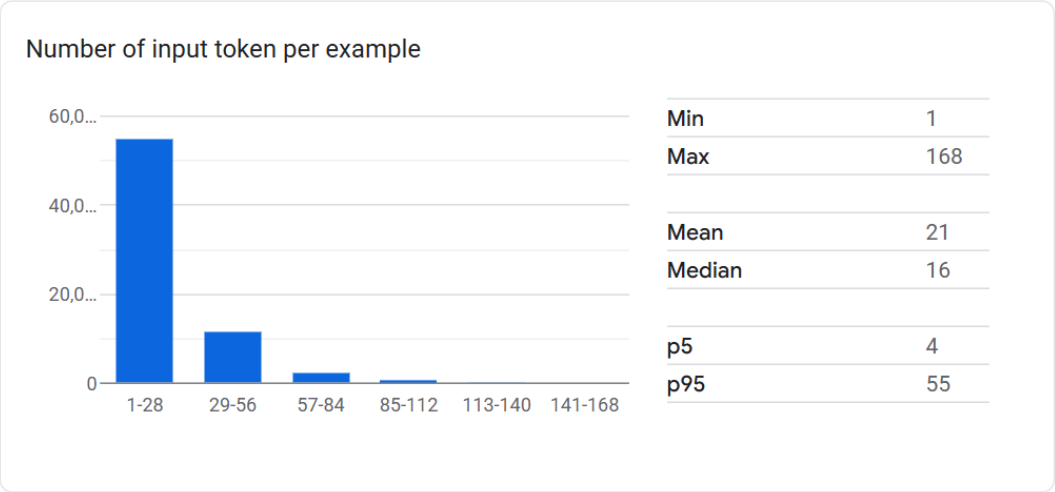


Figure 2. Number of input token per example

Figure 2 shows the distribution of the number of input tokens per example after fine-tuning, indicating that most of the data consists of short text. The number of tokens in the dataset varies, with a minimum of 1, a maximum of 168, an average of 21, and a median of 16 tokens. In addition, the p5 and p95 percentiles are 4 and 55 tokens, respectively, showing that the majority of examples fall within this range.

The majority of the text samples have between 1–28 tokens. The range of 29–56 tokens includes a smaller but still significant portion of the dataset. The number of examples decreases significantly as the token count increases, with very few examples containing more than 100 tokens, indicating their rarity. This distribution confirms that the fine-tuning dataset is dominated by short to medium-length text, with only a small portion consisting of long form content. The next data distribution, the number of output tokens per example, is shown in the following figure:

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

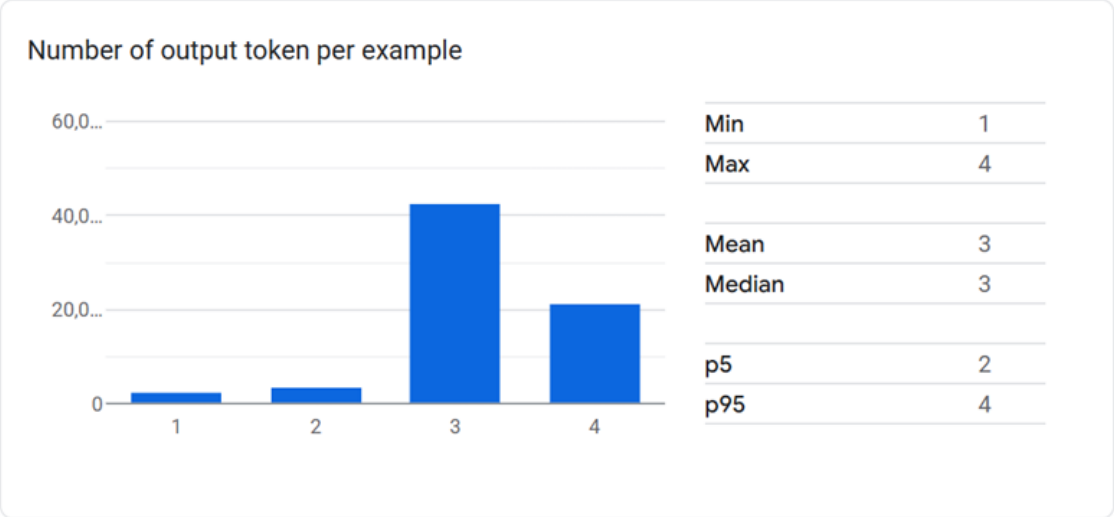


Figure 3. Number of output token per example

Figure 3 shows the distribution of the number of output tokens per example after fine-tuning. The model consistently produces short outputs, with token lengths ranging from 1 to 4. The statistical summary indicates a mean and median of 3 tokens, with p5 at 2 tokens and p95 at 4 tokens, reflecting a narrow and stable output range. Most of the examples had an output length of 3 tokens, followed by a significant portion with 4 tokens, while only a small number of examples had 1 or 2 tokens. This distribution confirms that the model has been optimized for text classification or label-based prediction tasks that require short, concise out puts, such as perception or sentiment classification. The model is clearly not designed for long text generation, given the highly limited number of tokens per output. The next data distribution, the number of messages per example, can be seen in the figure below:

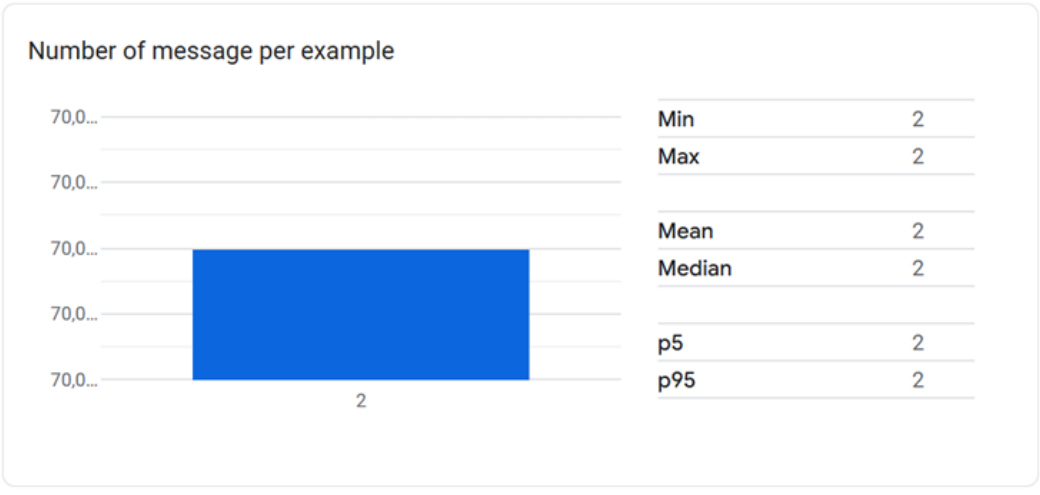


Figure 4. Number of message per example

Figure 4 shows the distribution of the number of messages per instance after fine-tuning, each example in the dataset has a fixed number of two messages. All statistics, including the minimum, maximum, mean, median, and p5 and p95 percentiles, are two, which confirms that there is no variation in the number of messages. The data visualization also

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

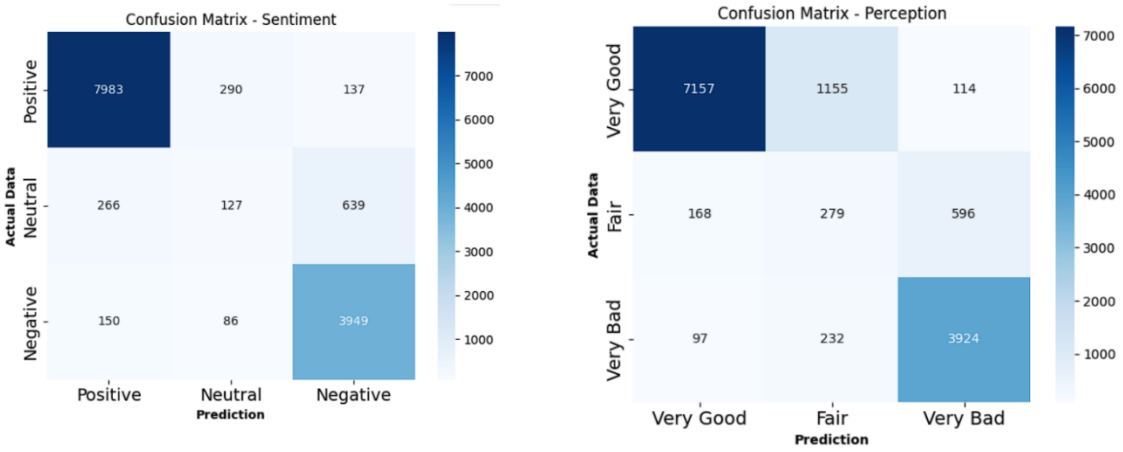
shows only one bar in the figure, which indicates that the structure of the dataset is very uniform. This suggests that the dataset likely contains short reviews or responses to a service, such as user comments on app features or complaints about technical issues. If the model is to be applied to analyze longer reviews, a more varied dataset is required. After the fine-tuning process and data distribution analysis were completed, the model was implemented on the prepared dataset. The next stage is testing with new data to ensure that the classification results are as expected.

The first evaluation is done by measuring Accuracy, Precision, Recall, and F1 Score, which are the main metrics in sentiment analysis. The evaluation results of the model are shown in the following table:

Tabel 3. Evaluation Metrics

Evaluation Metrics	Accuracy	Precision	Recall	F1 Score
Perception Result	0.6341	0.3298	0.2596	0.2882
Sentiment Result	0.6731	0.3398	0.2541	0.2875

The evaluation results of the fine-tuned Gemini 1.5 Flash model on the testing dataset consisting of 17,916 user reviews yielded an accuracy of 63.41% for perception classification and 67.31% for sentiment classification. The precision values obtained were 32.98% for perception and 33.98% for sentiment, while the recall values were 25.96% and 25.41%, respectively. The F1-scores reached 28.82% for perception and 28.75% for sentiment. These results indicate that the model was able to classify the majority of the data correctly with moderate performance across all metrics in both classification tasks. perception distribution and sentiment distribution are shown in the following graphs and matrices:



(a) (b)

Figure 5. Confusion Matrix (a) Confusion Matrix Sentiment; (b) Confusion Matrix Perception

The visualization results on the confusion matrix show that as many as 7,157 reviews indicate very good perception, 279 reviews indicate fair perception, and 3,924 reviews indicate very bad perception. Thus, in the perception distribution, the majority of reviews fall into the very good category, indicating a dominant perception of satisfaction among users

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

toward the BSI Mobile application. The very bad perception accounts for a significant portion as well, while the fair perception category represents the smallest portion of the reviews.

In terms of sentiment, 7,983 reviews fall into the positive category, 127 reviews are categorized as neutral, and 3,949 reviews are categorized as negative. This shows that the sentiment distribution is predominantly positive, with more than 60% of the reviews expressing satisfaction. The negative sentiment, which is close to 30%, reflects user concerns and dissatisfaction, whereas neutral sentiment remains minimal, indicating reviews that are more informative and less emotionally charged.

Overall, the distribution suggests that the BSI Mobile app continues to receive largely positive feedback, with positive reviews significantly outnumbering both neutral and negative ones. The high proportion of very good perception and positive sentiment demonstrates strong user approval, although a notable fraction of critical reviews still exists.

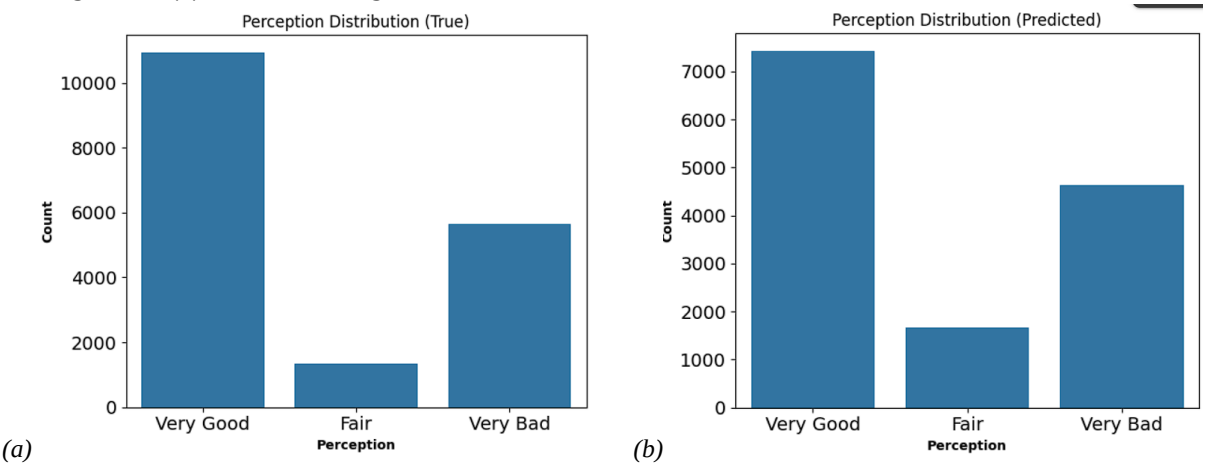


Figure 6. Chart Perception (a) Perception Distribution True; (b) Perception Distribution Predicted

This figure presents the original perception distribution derived from the *IndoBERT* based preprocessing stage before model fine-tuning. The dataset shows a dominant portion of reviews labeled as "Very Good", amounting to over 11,000 instances. This is followed by approximately 5,500 reviews categorized as "Very Bad", and a much smaller portion around 1,300 reviews classified as "Fair". This distribution reveals that the dataset used for fine-tuning was largely composed of polarized user experiences, with most users expressing either strong satisfaction or strong dissatisfaction. The relatively low occurrence of the fair category suggests that users tend to express clear-cut opinions regarding the BSI Mobile application.

This figure illustrates the predicted perception distribution after classification using the Gemini 1.5 Flash model. The majority of the reviews over 7,000 are classified as having a "Very Good" perception, which signifies a strong overall approval of the BSI Mobile application by users. The "Very Bad" category follows with around 4,000 reviews, indicating a significant minority of highly dissatisfied users. This distribution aligns with the previous sentiment analysis results and highlights the model's ability to distinguish different levels of user perception accurately.

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

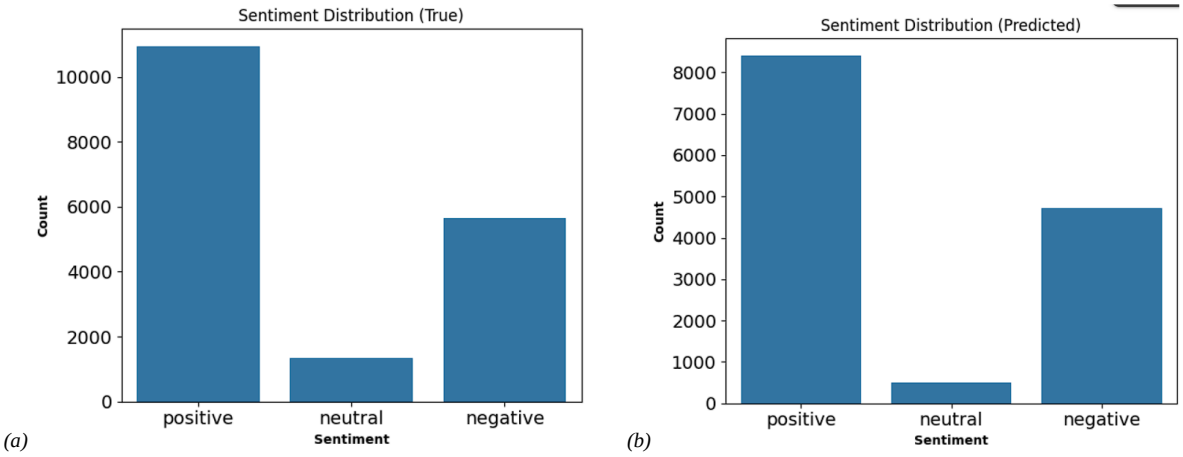


Figure 7. Chart Sentiment (a) Sentiment Distribution True; (b) Sentiment Distribution Predicted

This chart presents the ground truth sentiment distribution, which reflects the labels generated during the *IndoBERT* based preprocessing stage. Similar to the predicted sentiment, the true sentiment is also skewed towards the positive category, with more than 10,000 reviews labeled as positive. Approximately 5,500 reviews fall under the negative sentiment category, while around 1,500 reviews are labeled as neutral. The consistency between the predicted and true distributions reinforces the reliability of the fine-tuned model and the quality of the preprocessing pipeline. It also indicates that the *IndoBERT* based labeling process was effective in capturing the sentiment tendencies embedded in the Indonesian-language app reviews.

The predicted sentiment distribution graph shows a clear dominance of positive sentiment in the BSI Mobile user reviews. Out of the evaluated samples, over 8,000 reviews are categorized as positive, followed by nearly 5,000 reviews marked as negative, and only a small fraction falling into the neutral category. This distribution aligns with the confusion matrix and confirms that the majority of users express satisfaction with the application. The relatively high count of negative reviews also emphasizes the importance of addressing user concerns, even in the presence of a generally positive sentiment trend. The low volume of neutral sentiment may reflect a user tendency to express clear opinions either satisfaction or dissatisfaction rather than balanced or neutral feedback. The results of the manual evaluation with text are shown below:

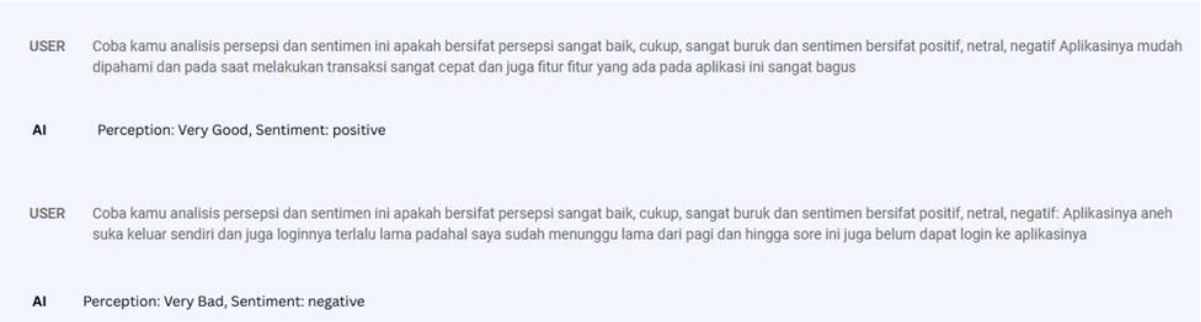


Figure 8. Manual Testing

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

Manual evaluation results through the chat feature in *Google Cloud Vertex AI* show that the model is able to classify perceptions and sentiments well based on the context of user sentences. In the first review, the model identified a positive sentiment corresponding to the user's satisfaction with the features and fast transactions. In the second review, the model identified a negative sentiment, centring on a complaint about the system's sluggish activation of the app. This result underscores the model's ability to recognize keywords effectively and understand context in perception and sentiment analysis. Nonetheless, additional testing is necessary to ensure its accuracy when dealing with ambiguous reviews or more nuanced opinions.

Algorithm 1

Review rating consistency evaluation

```
function check_consistency(rating, perception)
  if rating ≤ 2 and perception == "Sangat Buruk" then
    return "Consistent: Rating ≤2 & Very Bad"
  else if rating ≥ 4 and perception == "Sangat Baik" then
    return "Consistent: Rating ≥4 & Very Good"
  else if rating == 3 and perception == "Cukup" then
    return "Consistent: Rating = 3 & Fair"
  else if rating == 3 and perception == "Sangat Buruk" then
    return "Inconsistent: Rating = 3 & Very Bad"
  else if rating == 3 and perception == "Sangat Baik" then
    return "Inconsistent: Rating = 3 & Very Good"
  else
    return "Inconsistent: Others"
  end if
end function
```

Meanwhile, an analysis of the consistency and inconsistency of BSI Mobile user reviews was conducted, which in this analysis was taken from the perception of each review compared to the rating of the review, where if the review is positive but the rating is less than three stars, the review is inconsistent and vice versa, but if the review is positive and the rating is given 4 or 5 stars, the review is consistent and vice versa. Evaluation results of consistent and inconsistent reviews:

Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)
Volume 4 , Number 05, 2025, DOI 10.58471/esaprom.v4i05
ESSN 2828-9463 (Online)
https://ejournal.seaninstitute.or.id/index.php/esaprom

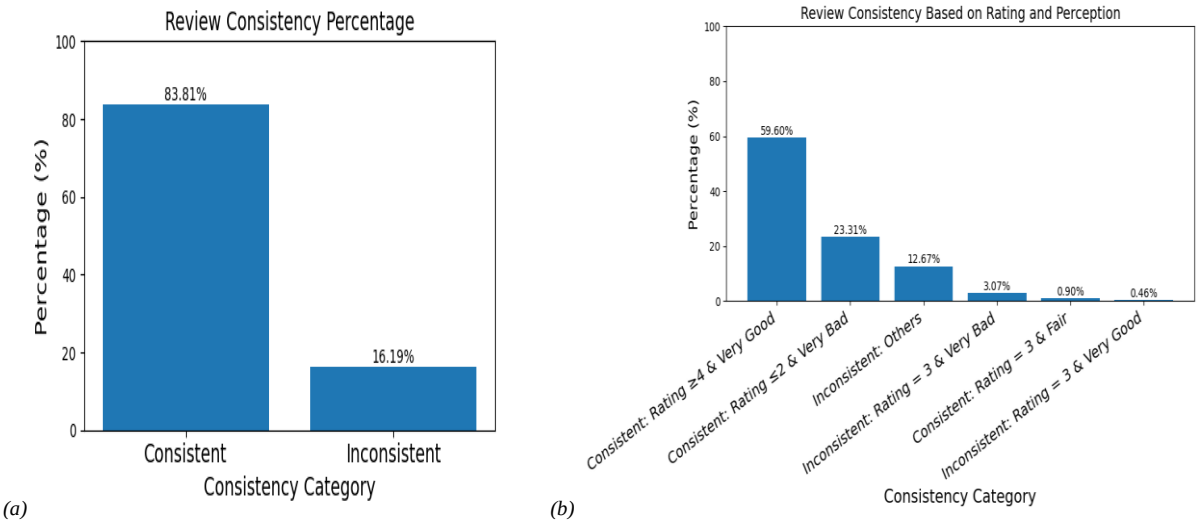


Figure 9. Percentage Of Inconsistency (a) Review Consistency Percentage; (b) Consistency Percentage Based On Rating And Perception

Figure 9 shows the distribution between consistent and inconsistent reviews of BSI Mobile app user ratings. The analysis shows that the majority of reviews fall into the consistent category, 83.81%, while inconsistent reviews amount to 16.19%. Furthermore, Figure 8 shows the details of consistency by rating and perception category. Consistent reviews with a rating of 4-5 and a perception of Very Good have the highest proportion, at 59.60%. Followed by reviews with a rating of 1-2 and a perception of Very Poor at 23.31%, and other consistent categories at 12.67%. Meanwhile, the inconsistent category consists of a combination of reviews with unaligned ratings and perceptions, namely: rating 1-2 and perception of Very Good by 3.07%, rating 3 and perception of Fair by 0.90%, and rating 3 and perception of Very Good by 0.46%.

The fine-tuned Gemini 1.5 Flash model demonstrated moderate classification performance in identifying user perception and sentiment from BSI Mobile reviews, with accuracy scores of 63.41% and 67.31% respectively. Although these results indicate the model's ability to capture general user sentiment, relatively low precision, recall, and F1-scores 25–34% suggest limitations in detecting nuanced or less frequent patterns, such as mixed or ambiguous feedback. One contributing factor may be the reliance on automatically labeled data using *IndoBERT*, which while effective in leveraging Indonesian language context can still introduce labeling noise or bias, particularly in edge cases like sarcasm or inconsistent expressions. Additionally, the dataset was highly polarized, with a large portion of reviews classified as “Very Good” or “Very Bad” and only a small fraction labeled as “Fair” leading to class imbalance that likely impacted the model’s learning and prediction accuracy.

Despite these limitations, the use of Gemini 1.5 Flash remains promising due to its scalability and adaptability for NLP tasks involving Indonesian Language, especially when combined with domain-specific fine-tuning. Compared to traditional models such as *Naïve Bayes* or SVM, which often show high accuracy on smaller or balanced datasets, LLMs offer greater contextual depth and robustness on large, unstructured corpora like app reviews. Moving forward, performance could be improved through manual validation of labels,

augmentation of minority classes, and the inclusion of metadata for contextual enrichment. Future studies should also assess the model's generalizability across other fintech platforms and explore its integration into real-time monitoring tools to support digital banking service improvement.

CONCLUSION

This study concludes that fine-tuning the Gemini 1.5 Flash model using a well-preprocessed and labeled dataset of 120,000 BSI Mobile reviews provides a moderately effective approach for classifying user perceptions in Indonesian-language digital banking contexts. With classification accuracy reaching 63.41% for perception and 67.31% for sentiment, the model showed potential in capturing general sentiment trends, despite limitations in precision and recall due to label noise and class imbalance. These results affirm the feasibility of applying large language models for local user feedback analysis, though further improvements are needed in data quality and model robustness to enhance practical application.

REFERENCES

Ahmad, K. *et al.* (2022) '*Global User-Level Perception of COVID-19 Contact Tracing Applications: Data-Driven Approach Using Natural Language Processing*', *JMIR Formative Research*, 6(5), p. e36238. Available at: https://doi.org/10.2196/36238.

Al-Baity, H.H. *et al.* (2022) '*Computational Linguistics Based Emotion Detection and Classification Model on Social Networking Data*', *Applied Sciences*, 12(19), p. 9680. Available at: https://doi.org/10.3390/app12199680.

Arifiyanti, A.A., Shantika, N.R. and Syafira, A.O. (2023) '*ANALISIS SENTIMEN ULASAN PENGGUNA BSI MOBILE PADA GOOGLE PLAY DENGAN PENDEKATAN SUPERVISED LEARNING*', *Jurnal Informatika Polinema*, 9(3), pp. 283–288. Available at: https://doi.org/10.33795/jip.v9i3.1003.

Arora, P. and Banerji, R. (2024) '*The impact of digital banking service quality on customer loyalty: An interplay between customer experience and customer satisfaction*', *Asian Economic and Financial Review*, 14(9), pp. 712–733. Available at: https://doi.org/10.55493/5002.v14i9.5199.

Beurer-Kellner, L., Fischer, M. and Vechev, M. (2023) '*Prompting Is Programming: A Query Language for Large Language Models*', *Proceedings of the ACM on Programming Languages*, 7(PLDI), pp. 1946–1969. Available at: https://doi.org/10.1145/3591300.

Catania, C. *et al.* (2022) '*Beyond Random Split for Assessing Statistical Model Performance*', pp. 1–12. Available at: https://typeset.io/papers/beyond-random-split-for-assessing-statistical-model-2bxbb9rd.

Fahrani, F. and Aryanto, J. (2024) '*Sentiment Analysis of Public Opinion on the Palestinian-Israeli Conflict using Support Vector Machine and Naïve Bayes Algorithms*', *Journal of Scientific Research, Education, and Technology (JSRET)*, 3(4), pp. 1890–1900. Available at: https://doi.org/10.58526/jsret.v3i4.606.

Fatouros, G. *et al.* (2024) '*Can Large Language Models beat wall street? Evaluating GPT-4's*

impact on financial decision-making with MarketSenseAI', *Neural Computing and Applications* [Preprint]. Available at: https://doi.org/10.1007/s00521-024-10613-4.

Gerlich, M., Elsayed, W. and Sokolovskiy, K. (2023) '*Artificial intelligence as toolset for analysis of public opinion and social interaction in marketing: identification of micro and nano influencers*', *Frontiers in Communication*, 8. Available at: https://doi.org/10.3389/fcomm.2023.1075654.

Kaur, H. and Sandhu, N.K. (2023) '*International Journal of Communication Networks and Information Security Evaluating the Effectiveness of the Proposed System Using F1 Score , Recall , Accuracy , Precision and Loss Metrics Compared to Prior Techniques*', 15(04), pp. 368–383.

Liu, Y. *et al.* (2023) '*Improving Large Language Model Fine-tuning for Solving Math Problems*', (1), pp. 1–14. Available at: http://arxiv.org/abs/2310.10047.

Malik, N. and Bilal, M. (2024) '*Natural language processing for analyzing online customer reviews: a survey, taxonomy, and open research challenges*', *PeerJ Computer Science*, 10, p. e2203. Available at: https://doi.org/10.7717/peerj-cs.2203.

Olujimi, P.A. and Ade-Ibijola, A. (2023) '*NLP techniques for automating responses to customer queries: a systematic review*', *Discover Artificial Intelligence*, 3(1), p. 20. Available at: https://doi.org/10.1007/s44163-023-00065-5.

Sodik, F., Nur Zaida, A. and Zulmiati, K. (2022) '*Analisis Minat Penggunaan pada Fitur Pembelian Mobile Banking BSI: Pendekatan TAM dan TPB*', *Journal of Business Management and Islamic Banking*, 1(1), pp. 35–53. Available at: https://doi.org/10.14421/jbmib.2022.011-03.

Sottana, A. *et al.* (2023) '*Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks*', in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 8776–8788. Available at: https://doi.org/10.18653/v1/2023.emnlp-main.543.

Trani, A.H. and Tran, D.A. (2024) '*CUSTOMER EXPERIENCE AND SATISFACTION WITH DIGITAL BANKING SERVICES*', *Proceeding of International Conference on Business, Economics, Social Sciences, and Humanities*, 7, pp. 548–555. Available at: https://doi.org/10.34010/icobest.v7i.565.

Wong, M.-F. *et al.* (2023) '*Natural Language Generation and Understanding of Big Code for AI-Assisted Programming: A Review*', *Entropy*, 25(6), p. 888. Available at: https://doi.org/10.3390/e25060888.