Artificial Intelligence Systems and Its Applications (AISA)
Vol. 1, No. 1, July 2025
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

SENTIMENT ANALYSIS OF PUBLIC FEEDBACK ON UNIVERSITAS MUHAMMADIYAH SURAKARTA THROUGH GOOGLE MAPS REVIEWS: INSIGHTS AND IMPLICATIONS

***Agus Ardiansyah Nh*[1], *Endang Wahyu Pamungkas*[2], *Sohail Akhtar*[3]**

[1,2]Department of Informatics Engineering, Faculty Communication and Informatics,
Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

[3]Department of Computer Science, Faculty of Engineering Science, Bahria University,
Islamabad, Pakistan

Abstract: In the era of digital feedback, higher education institutions face growing challenges in making sense of large volumes of user-generated reviews, especially those written in multiple languages. This study analyzes 1,717 Google Maps reviews related to Universitas Muhammadiyah Surakarta (UMS), collected over five years in Bahasa Indonesia and English. To overcome limitations of manual and monolingual sentiment analysis, we employed a pre-trained multilingual transformer model—lxyuan/distilbert-base-multilingual-cased-sentiments-student—without additional fine-tuning. The analysis revealed that 88% of reviews were classified as positive, with most praise directed at campus facilities, while criticism often targeted administrative services. Beyond sentiment classification, this study explored text length, confidence scores, and user engagement patterns to uncover deeper behavioral insights. We also developed SentiMu, an interactive dashboard that visualizes sentiment trends, recent reviews, word clouds, and key metrics, enabling university stakeholders to monitor feedback in real time. The dashboard was built using Next.js and FastAPI for optimal performance and scalability. By automating the analysis and visualization of multilingual online reviews, this study provides a practical and scalable framework for institutions to understand student and visitor experiences, supporting data-driven decisions to enhance campus services and reputation.

Keywords: Google Maps Review, Interactive Data Visualization, Multilingual Sentiment Analysis, Satu Data, Sentiment Analysis.

Article History:

Received: 25 April 2025 Accepted: 28 July 2025 Published: 28 July 2025

Corresponding Author: Agus Ardiansyah Nh, Email: aguzardiansyahnh@gmail.com

DOI: 10.65917/aisa.v1i1.12

1. Introduction

In the digital era, higher education institutions are increasingly challenged to make sense of the growing volume of user-generated feedback available on platforms such as Google Maps [1]. Two major barriers often hinder this process. First, the overwhelming number of reviews makes it difficult to analyze feedback efficiently, often leading to missed insights. Second, the multilingual nature of reviews particularly in diverse academic environments complicates the accurate interpretation of user sentiment expressed in different languages.

Traditionally, universities used tools like surveys and interviews conducted at term's end, which were limited in scope and timeliness. The rise of platforms like Google Maps has shifted this landscape, enabling real-time, unsolicited feedback that offers broader, dynamic insights into campus experiences and allows institutions to respond more proactively [2]. Additionally, online reviews increasingly influence prospective students' decision-making processes, making the effective analysis of this feedback more crucial than ever.

Several studies have begun to explore the potential of Google Maps reviews in educational contexts. For instance, [3] analyzed library user feedback using open coding of Google Maps reviews, showing how such reviews can capture meaningful user experiences. Matthews [4] further investigated how Google Maps reviews reflect perceptions of public library spaces, emphasizing how user language and tone contribute to perceived value. Similarly, Shoemaker [5] conducted a thematic analysis of public library reviews using Google Maps, categorizing them into spatial, emotional, and service-based feedback. These studies demonstrate the utility of Google Maps data in understanding user experiences but rely heavily on manual coding.

While valuable, these manual approaches limit scalability, often capping analysis at around 500 reviews. Most studies also focus on monolingual data, missing the complexity and richness of multilingual sentiment [6]. This creates a gap in research methodology, especially when applied to settings like Indonesian universities where multiple languages are used in daily communication [7].

To address the limitations of manual and monolingual analysis, this study adopts a transformer-based approach using BERT to analyze approximately 2,000 Google Maps reviews written in Bahasa Indonesia and English, collected over a five-year period [8]. This multilingual dataset enables the identification of sentiment trends over time, while an interactive dashboard allows stakeholders to explore keyword-based sentiment across categories such as academic programs, administrative services, and campus facilities.

This research aims to develop a multilingual sentiment analysis pipeline integrated with visual analytics. It also seeks to uncover temporal sentiment patterns and relate them to institutional events such as renovations or policy shifts. Furthermore, the study explores how sentiment differs across various aspects of campus life, based on the hypothesis that physical facilities may generate more positive sentiment compared to administrative services. The novelty of this research lies in its direct application of transformer-based multilingual models to Google Maps reviews of a higher education institution in Indonesia, without requiring fine-tuning, and its integration into a real-time interactive dashboard for institutional monitoring. Ultimately, the goal is to offer a scalable, data-driven system that enhances institutional responsiveness and improves overall campus experience for students and visitors. The next section reviews related studies in higher education sentiment analysis and outlines the need for an automated, multilingual framework.

2. Related Work (if applicable)

Previous studies have provided valuable insights through manual coding, enabling in-depth qualitative analysis of user-generated reviews. However, their scalability remains limited, often constrained to fewer than 500 monolingual reviews, which restricts the identification of broader patterns and long-term sentiment trends. Moreover, these approaches typically overlook the complexity of multilingual feedback that reflects the diverse backgrounds of university stakeholders. To overcome these limitations, this study applies automated sentiment analysis techniques to process thousands of multilingual reviews, allowing for a more comprehensive and scalable understanding of user sentiment. By doing so, the research enhances both the depth and generalizability of insights while building upon the foundational work established by earlier studies [6]. This section has outlined prior research and highlighted existing limitations in sentiment analysis methods, particularly for multilingual data. To address these gaps, the following section introduces the methodological framework designed to support the implementation and evaluation of multilingual transformer-based sentiment analysis models.

3. Methods

3.1 Research Design

This study uses an experimental comparative approach to assess how well different multilingual transformer-based models perform in classifying sentiments from Google Maps reviews. The process includes collecting and cleaning review data, training and fine-tuning selected models, and evaluating their performance using common metrics such as accuracy and F1-score. As illustrated in Figure 1, the research process begins with collecting publicly available reviews from Universitas Muhammadiyah Surakarta on Google Maps. The figure outlines the key steps in the methodology, from data acquisition and preprocessing to model development and evaluation.

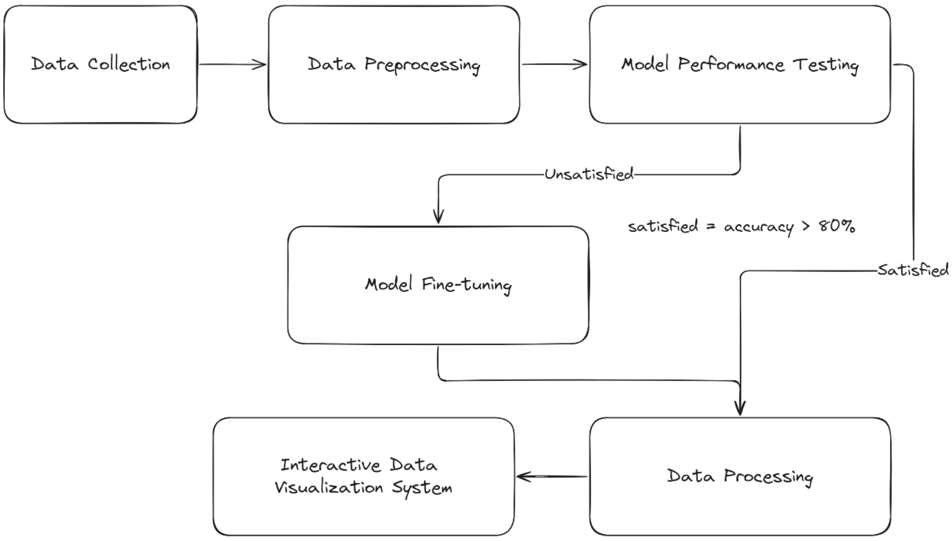

Figure 1 : Research Methodology Workflow.

3.2 Data Collection

The data collection process used a web scraping method with the Apify platform, specifically the Google Maps Review Scraper actor, to collect targeted reviews [9]. This tool was set up with criteria such as multiple location URLs of Universitas Muhammadiyah Surakarta, sorted by the latest reviews, and no review limits. The scraping covered reviews from January 2019 to January 2025 to ensure relevance. To support model evaluation, we also gathered reviews from a nearby university, which will be explained further. Figure 2 shows the scraping process in detail.

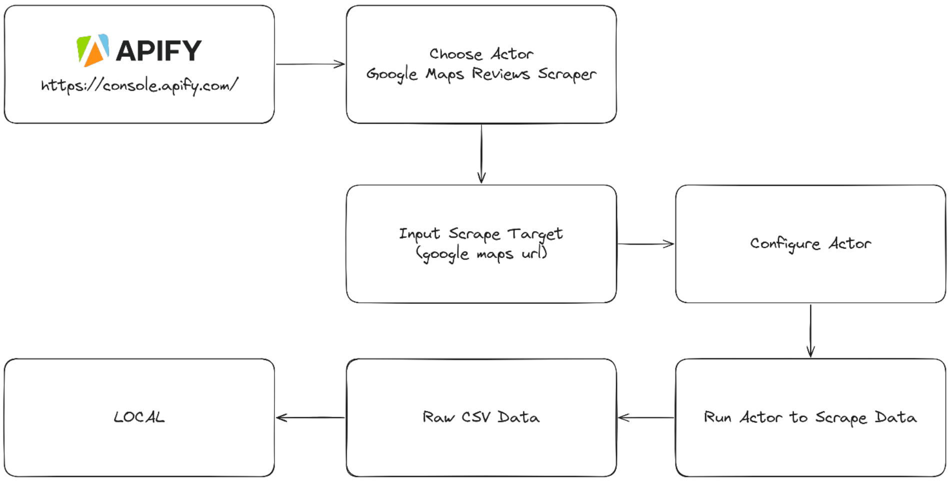

Figure 2 : Data Scraping Methodology.

3.3 Data Preprocessing

The data collected from the scraping process will undergo preprocessing to remove unnecessary information from the Google Maps reviews for a more structured dataset [10].

3.3.1 Cleaning

The cleaning process removes irrelevant or incomplete data, such as reviews with missing values, empty columns, or unnecessary identifiers [11]. The retained fields include:

1. publishedAtDate: For tracking the publication date of the review.
2. stars: Representing the rating score.
3. name: Reviewer's name.
4. reviewerId: For tracking user's activity
5. likesCount: The number of likes on the reviewer's review.
6. reviewerPhotoUrl: Reviewer's profile image.
7. reviewUrl: Providing a direct link to the review.
8. text: Containing the review itself.

3.3.2 Normalization

Normalization standardizes the review text while keeping its meaning intact. This involves removing unnecessary elements like HTML tags, extra punctuation, and emojis. Spelling mistakes and informal words (e.g., "gr8" to "great", "yg" to "yang") are corrected to standard forms [12]. However, uppercase and lowercase letters are kept as-is because our model is case-sensitive and uses that distinction to better understand the text.

3.3.3 Exploratory Data Analysis (Initial Data Exploration)

Exploratory data analysis was carried out to understand how sentiment ratings were distributed, check the data quality after cleaning, and see how reviews changed over time during the study period. The initial results showed how star ratings relate to sentiment labels, helping to uncover patterns and possible biases in the dataset [13].

For text preprocessing, we deliberately avoid traditional NLP steps such as stopword removal, manual tokenization, and stemming/lemmatization. This is because transformer models from Hugging Face already have advanced tokenizers that can analyze text at both the word and subword level. As demonstrated by [14] and [15], these models utilize specialized tokenizers that decompose text into contextually meaningful subword units. For instance, words like "running," "runs," and "runner" are broken down into related subword parts, allowing the model to understand their meaning more effectively. Applying traditional preprocessing manually could interfere with this built-in process and reduce model performance.

3.4 Model Performance Testing

3.4.1 Model Selection

There are numerous open-source platforms for sharing pre-trained models, with Hugging Face being one of the most prominent [16]. Given the vast array of pre-trained models available on Hugging Face, we curated our selection to align with our research focus on multilingual text classification models. Our investigation identified three models with the highest monthly active usage per October 2024:

1) **lxyuan/distilbert-base-multilingual-cased-sentiments-student (2.05M)**

   The lxyuan/distilbert-base-multilingual-cased-sentiments-student is a lightweight transformer model for multilingual sentiment analysis, supporting languages like English, Indonesian, and Japanese. Trained via zero-shot distillation from a larger teacher model (mDeBERTa-v3-base-mnli-xnli), it achieved an 88.29% agreement score. While efficient and reliable, the model may struggle with sarcasm, irony, or unfamiliar text styles [17].

2) **crypter70/IndoBERT-Sentiment-Analysis (9.49K)**

   This model is a fine-tuned version of indobenchmark/indobert-base-p1, optimized for sentiment analysis in Bahasa Indonesia using the IndoNLU dataset. It achieved 94.52% accuracy and a 94.51% F1-score, indicating high effectiveness in classifying Indonesian texts into positive, neutral, or negative sentiments [18].

3) **carant-ai/xlm-roberta-sentiment-base (243)**

   This model is a fine-tuned version of xlm-roberta-base, designed for sentiment analysis in both English and Bahasa Indonesia. Trained on labeled datasets in both languages, it can classify text into positive, neutral, or negative sentiment. Its bilingual capability makes it suitable for analyzing multilingual feedback, such as user reviews from diverse language backgrounds. [19].

3.4.2 Performance Testing

To identify the most suitable model for our sentiment analysis task, we conducted a comparative evaluation of the three selected models using standard multilingual sentiment classification metrics. For this purpose, we curated a dataset of 1,338 Google Maps reviews collected from a neighboring university. We applied an extreme review sampling method, a common technique in sentiment analysis that selects reviews from the two ends of the rating scale [20] :

1. Positive sentiment: 5-star ratings
2. Negative sentiment: 1-star ratings

This binary classification setup provides a clear ground truth by focusing on reviews with strong sentiment polarity [21]. It is particularly effective during the initial testing phase, enabling a reliable performance comparison before addressing more ambiguous or mixed-sentiment cases.

3.4.3 Model Evaluation and Fine-tuning

Following the performance testing, we will proceed with one of two approaches:

1. If one of the tested models meets our performance criteria which is accuracy more than 80%, we will select it for our research.

2. If none of the models sufficiently meet our needs, we will choose the model with the highest performance. This model will then undergo fine-tuning using our specially curated dataset we mention above.

Our dataset is designed to represent sentiment analysis in two specific contexts: Google Maps reviews and educational environments. The fine-tuning process aims to optimize the chosen model's performance for our specific use case, ensuring it is well-adapted to analyze sentiment in these particular area [22].

3.5 Model Fine-tuning

Our fine-tuning process leverages the Transformers library provided by Hugging Face, which offers a robust framework for adapting pre-trained models to specific tasks. We utilized the Trainer API, a built-in functionality that streamlines the fine-tuning process.

3.5.1 Data Preparation

1. Dataset Splitting
   The curated dataset of 1338 polarized Google Maps reviews was split into training and testing sets. We employed an 80:20 split ratio.
2. Text Preprocessing and Tokenization
   Input texts were tokenized using the model-specific tokenizer. Special tokens ([MASK]) were automatically handled by the tokenizer.

3.5.2 Fine-tuning Configuration

1. Training Parameters
   a. Output directory: ./results
   b. Evaluation strategy: epoch
   c. Learning rate: 2e-5 with linear decay
   d. Per device training batch size: 8
   e. Per device evaluation batch size: 8
   f. Number of training epochs: 3
   g. Weight decay: 0.01
2. Model Architecture
   a. Model: Assigned with chosen model.
   b. Tokenizer: Preserved from the pre-trained model.
   c. Model argument: Assined with training argument.
   d. Training dataset: Assigned with spliting our curated dataset.
   e. Eval dataset: Assigned with spliting our curated dataset.

3.5.3 Training Process

The fine-tuning process was implemented using the following steps:

```
training_args = TrainingArguments(
    output_dir="./results",
    eval_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
)
trainer = Trainer(
    model=model,
    tokenizer=tokenizer,
    args=training_args,
    train_dataset=tokenized_train,
    eval_dataset=tokenized_test
)
trainer.train()
```

3.5.4 Evaluation Metrics

1. Training loss

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \quad (1)$$

Training loss measures the difference between the model's predicted values and the actual outcomes during training. A steadily decreasing loss value usually indicates that the model is learning effectively and improving its predictions over time.

2. Validation accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Validation accuracy measures the overall correctness of the model's predictions on a separate validation dataset. It is calculated by dividing the number of correct predictions (both true positives and true negatives) by the total number of predictions. A higher accuracy indicates better general performance on unseen data.

3. Precision and recall

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

Precision answers: “Of all instances predicted as positive, how many are truly positive?” Recall answers: “Of all actual positive cases, how many did the model correctly identify?” Both are crucial for evaluating classification performance, especially in imbalanced datasets.

4. F1-score

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (5)$$

The F1-score is the harmonic mean of precision and recall. It balances both metrics into a single value, especially useful when there’s an uneven class distribution or when both false positives and false negatives matter.

5. Confusion matrix

Table 1 : Confusion Matrix

TP	FP
FN	TN

As shown in Table 1, the confusion matrix is a 2×2 table used in binary classification to present true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). It helps reveal not just how many errors a model makes, but also the types of misclassifications, offering insights into model performance. This study uses the scikit-learn library to compute evaluation metrics efficiently, including accuracy, precision, recall, F1-score, and the confusion matrix.

3.6 Data Processing

This phase involves systematically processing the cleaned dataset using advanced NLP techniques through the selected or fine-tuned transformer model. The goal is to ensure accurate text representation while preserving key semantic information for reliable sentiment classification.

3.6.1 Sentiment Analysis

The sentiment analysis was conducted using the selected or fine-tuned model, which is based on a transformer architecture specifically trained for sentiment classification tasks. The implementation included the following steps:

1. Model Initialization

The pre-trained model was loaded, and its parameters were configured to ensure optimal performance

2. Text Processing

Input texts were tokenized using the model's native tokenizer, with padding and truncation set to True to maintain uniform sequence lengths.

3. Sentiment Classification

The tokenized sequences were processed in batches and classified into predefined sentiment categories.

3.6.2 Explanatory Data Analysis (Post-Sentiment Analysis)

The post-sentiment analysis phase involved a comprehensive examination of the model output analysis through the following key aspects:

1. Distribution of predicted sentiment classifications

$$f = ni/N \quad (6)$$

*ni is count of each sentiment category, N is total samples

Analysis of sentiment category distributions across the dataset, including frequencies, and identification of dominant sentiment trends within the processed data.

2. Confidence scores distribution across predictions

$$\mu conf = (1/N)\sum ci \quad (7)$$

*ci is confidence score for each prediction

$$count(conf > threshold) = \sum I(ci>0.7) \quad (8)$$

*counts predictions where confidence score ci exceeds 0.7

A comprehensive assessment of the model’s prediction certainty was conducted through confidence threshold analysis. This approach evaluates the model’s reliability by quantifying the proportion of predictions made with high confidence (i.e., confidence scores greater than 0.7). By focusing on high-certainty outputs, this analysis supports a more robust and trustworthy sentiment classification.

3. Analysis of prediction patterns across different text lengths

$$r = cov(length, accuracy)/\sigma length \times \sigma accuracy \ (9)$$

*measures relationship between text length and accuracy

This analysis explores the relationship between text length and prediction accuracy, evaluating how input complexity affects model performance. It also aims to identify the optimal text length range that supports consistent and reliable sentiment classification.

3.6.3 Derived Feature Engineering

The feature engineering phase involved the systematic derivation of new features from the existing dataset to enhance the analytical capabilities:

1. Review Quality Indicators
   Review quality was assessed using several metrics. The length of reviews provided insights into the depth and detail of user feedback [23]. User engagement was analyzed by tracking posting frequency per user, helping identify highly active reviewers or potential bots. To preserve data integrity, spam likelihood indicators were also applied to flag suspicious or repetitive review patterns based on predefined criteria.
2. Temporal Features
   Temporal analysis examined chronological patterns in the review data. Monthly review counts were tracked to observe submission volumes over time. Long-term trends and growth trajectories were identified, along with seasonal peaks that may align with academic calendar events.
3. Content-Based Features
   Content-Based Features focused on identifying patterns within review texts. Common phrases were extracted to reveal recurring themes and standardized expressions, while keyword frequency analysis highlighted dominant terms and topics across the corpus, offering insights into the most discussed issues [24].

3.7 Interactive Data Visualization System

A user friendly web based dashboard was developed to display sentiment analysis results, using modern web technologies and efficient data processing to ensure smooth and interactive exploration of insights.

3.7.1 Dashboard Implementation

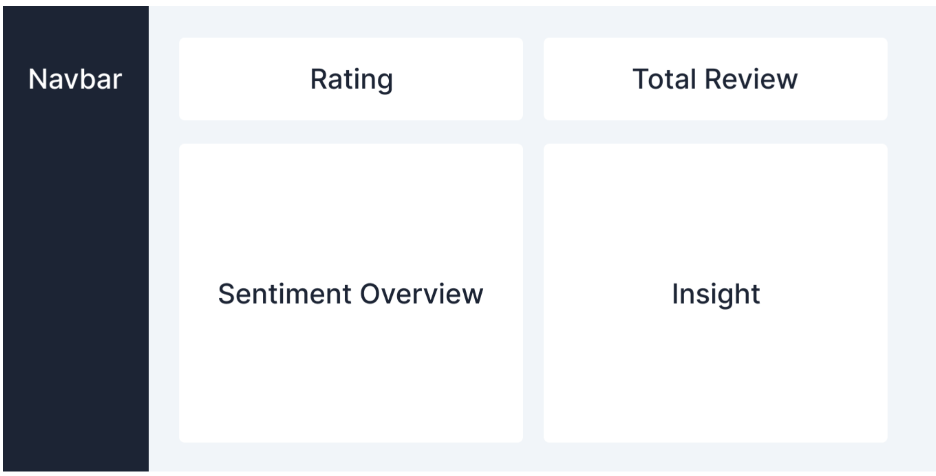

Figure : 3 Frontend Design.

As illustrated in Figure 3, the frontend of the dashboard is built using Next.js, TypeScript, and Tailwind CSS. This technology stack was selected to accommodate the system's complexity and to allow seamless integration of future features. Next.js, as one of the most widely used frameworks, supports server-side rendering and static site generation, offering enhanced performance suitable for scalable and maintainable web applications.

3.7.2 API Architecture

The backend was developed using FastAPI and Redis for caching. FastAPI was chosen because it is fast, supports asynchronous processing, and is easy to scale. Redis is used as a caching system to speed up data access and reduce the load on the main database. This tech stack ensures that the backend runs efficiently, responds quickly, and is easy to maintain. With the methodological framework firmly established covering data acquisition, preprocessing, model evaluation, and dashboard design this section proceeds to present the analytical results. The findings are discussed in detail to highlight sentiment trends, user behaviors, and the overall performance of the developed system.

4. Results and Discussion

This study began with a comprehensive evaluation of various pretrained models using a custom dataset collected from neighboring universities. This evaluation was essential to identify the most appropriate model for conducting sentiment analysis on Google Maps reviews pertaining to Universitas Muhammadiyah Surakarta. The performance testing revealed that the selected pretrained models met or exceeded our predetermined accuracy thresholds, eliminating the need for additional fine-tuning in this research.

4.1 Data Collection

The systematic collection and curation of Google Maps data yielded 25 distinct points associated with Universitas Muhammadiyah Surakarta (UMS). These locations represent various university facilities, buildings, and affiliated sites across the campus area. Table 2 presents a comprehensive overview of these geographical points.

Table 2 : Google Maps Point URL related to UMS

No	Url	Description
01	https://maps.app.goo.gl/STR6DFzamKqvdpL67	FKI UMS
02	https://maps.app.goo.gl/cgPzqCAxskRyQjZf7	FEB
03	https://maps.app.goo.gl/NnoMCZXHi8NzdXnt8	FT
04	https://maps.app.goo.gl/SHT1tb5oX19nDWH77	FIK
05	https://maps.app.goo.gl/94kqZgTjheGw3jVJ7	FAI
06	https://maps.app.goo.gl/epB4ycGwknAixsKJ8	Ffarmasi

07	https://maps.app.goo.gl/aQoQRaxdJbxvtzjw9	FKIP
08	https://maps.app.goo.gl/Qf52ujsfHxyg93oX7	FH
09	https://maps.app.goo.gl/kdQRK1fNUe1GP8hP6	Fgeo
10	https://maps.app.goo.gl/3PCsmNGetsSugq2o7	Fpsi
11	https://maps.app.goo.gl/g3CY3N98vJnike7P8	FKG
12	https://maps.app.goo.gl/E6rMPez5JfDUpENX6	FK
13	https://maps.app.goo.gl/6wYEjFVwYkKLzZTE6	Danau Salsabila
14	https://maps.app.goo.gl/KvWRBfrw3QjgUbZg6	Kantin Tepi Danau
15	https://maps.app.goo.gl/wNRL7RXvWkSDwGB99	Lapangan Tenis
16	https://maps.app.goo.gl/9MysxmgqxxZoLzFN7	GOR Kampus 2
17	https://maps.app.goo.gl/fPY1e2Dja4YiEjst9	Kampus 1
18	https://maps.app.goo.gl/iFGubshw2GktPMtL9	Kampus 4
19	https://maps.app.goo.gl/GVQHbFDBpsbs1aHY8	UMS
20	https://maps.app.goo.gl/ufV1pWNLitK3GjK58	Masjid Sudalmiyah
21	https://maps.app.goo.gl/QsshP9dRQTJetog18	Gedung Siti Walidah
22	https://maps.app.goo.gl/FWYobH9AFuy5QnBu8	Perpustakaan
23	https://maps.app.goo.gl/qzjQwi4bnha9md1w8	Auditorium Djazman
24	https://maps.app.goo.gl/jAoiwWFc2mY7uq1g9	Gedung Pasca Sarjana
25	https://maps.app.goo.gl/R7S5jF3k5aEPCWUK6	BKPP Psikologi

The data collection process was executed using the Apify platform's specialized actor for Google Maps reviews. The scraping operation was conducted on January 15, 2025, at 10:19 UTC+7, following the systematic compilation of relevant Google Maps URLs associated with Universitas Muhammadiyah Surakarta (UMS). This automated extraction method was designed to retrieve comprehensive review data from 25 identified UMS-related locations across different faculties and campus facilities. Figure 4 illustrates the result interface of the Apify actor, which documents the execution log and collected review entries. This visual output validates the successful retrieval process and helps confirm the scraping parameters used during the operation.

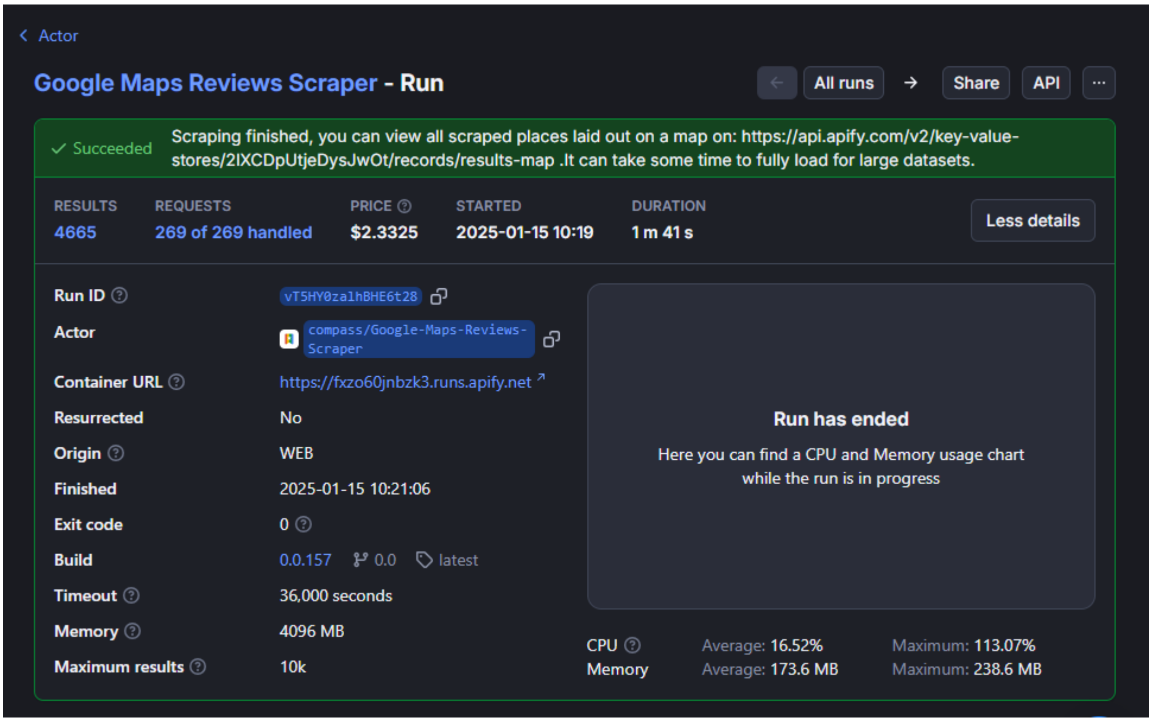

Figure : 4 Apify Actor Result.

The web scraping operation yielded the following outcomes:

Table 3 : Data Scrape Information

Information	Description
Number of reviews scrapped	4665 Reviews
Number of dataset columns	81 Columns
Newest published review	14 Jan 2025 at 23:02:05 UTC+7
Oldest published review	01 Jan 2019 at 01:02:58 UTC+7

As shown in Table 3, the scraping operation resulted in a total of 4,665 reviews across 81 structured columns. The temporal range of the collected data spans from the oldest review published on January 1, 2019, to the most recent review on January 14, 2025, indicating that the dataset captures a broad time frame that reflects evolving user sentiment across several academic periods.

4.2 Data Preprocessing

The data collected from the scraping process will undergo preprocessing to remove unnecessary information from the Google Maps reviews for a more structured dataset.

1. Cleaning

   The cleaning process focused on removing irrelevant or incomplete entries. Specifically, reviews with missing (NaN) values, empty columns, and unnecessary identifiers were removed, ensuring only essential data remained. As a result of this cleaning step, the dataset was reduced to 1,718 reviews and 7 key columns. Figure 5 illustrates the result of the dataset after the cleaning process, highlighting the reduced number of records and retained features.

```
Total dataset size before removing NaN: 4665
Total dataset size after removing NaN: 1718

========================================================================

<class 'pandas.core.frame.DataFrame'>
Index: 1718 entries, 0 to 4664
Data columns (total 7 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   publishedAtDate   1718 non-null   object
 1   stars             1718 non-null   int64
 2   likesCount        1718 non-null   int64
 3   name              1718 non-null   object
 4   reviewerPhotoUrl  1718 non-null   object
 5   reviewUrl         1718 non-null   object
 6   text              1718 non-null   object
dtypes: int64(2), object(5)
memory usage: 107.4+ KB
None
```

Figure : 5 Dataset Cleaning Result.

2. Normalization

To ensure that the text data is standardized while preserving its meaning, we applied a normalization process to the reviews. This step included:

- Removing non-informative elements such as HTML tags, excess punctuation, and emojis.
- Correcting spelling errors and informal abbreviations to improve text consistency.
- Maintaining case sensitivity to leverage a cased pretrained model that distinguishes between uppercase and lowercase letters.

The initial dataset consisted of 1,718 reviews collected from Google Maps. The Python script processed each review and applied the normalization techniques described in Chapter 2. Table 4 presents the dataset statistics before and after processing.

Table 4 : Dataset Normalization Result

Dataset State	Total Reviews
Initial dataset size	1718
After duplicate removal	1717
Duplicates removed	1 (0.06%)

The dataset contained a small number of duplicate reviews, which were identified and removed. The remaining 1,717 unique reviews proceeded to the next preprocessing steps. As part of the preprocessing pipeline, we detected the language of each review to assess its distribution. Table 5 shows the number of reviews per detected language.

Table 5 : Dataset Languange Distribution

Detected Language	Total Review	Percentage
Indonesia	1312	76.4%
English	90	5.2%
None/Unknown	315	18.4%

The majority of the dataset (76.4%) consists of Indonesian-language reviews, followed by 18.4% unidentified or mixed-language reviews, and 5.2% English reviews. This distribution indicates that sentiment analysis will primarily focus on the Indonesian language.

3. Exploratory Data Analysis (Initial Data Exploration)

The analysis of 1,717 Google Maps reviews revealed a clear pattern in language distribution. The majority of reviews were written in Indonesian, totaling 1,312 entries (76.4%), indicating strong engagement from local users. English reviews accounted for 90 entries (5.2%), reflecting a smaller international presence. Meanwhile, 315 reviews (18.4%) were categorized as None or Unknown language due to inconsistencies, code-switching, or the use of unsupported languages. This distribution highlights that Universitas Muhammadiyah Surakarta primarily receives feedback from an Indonesian-speaking audience, while also gaining attention from a global user base. Figure 6 presents the distribution of star ratings across each language group, offering insight into how sentiment represented by rating scores varies depending on the language used in the reviews.

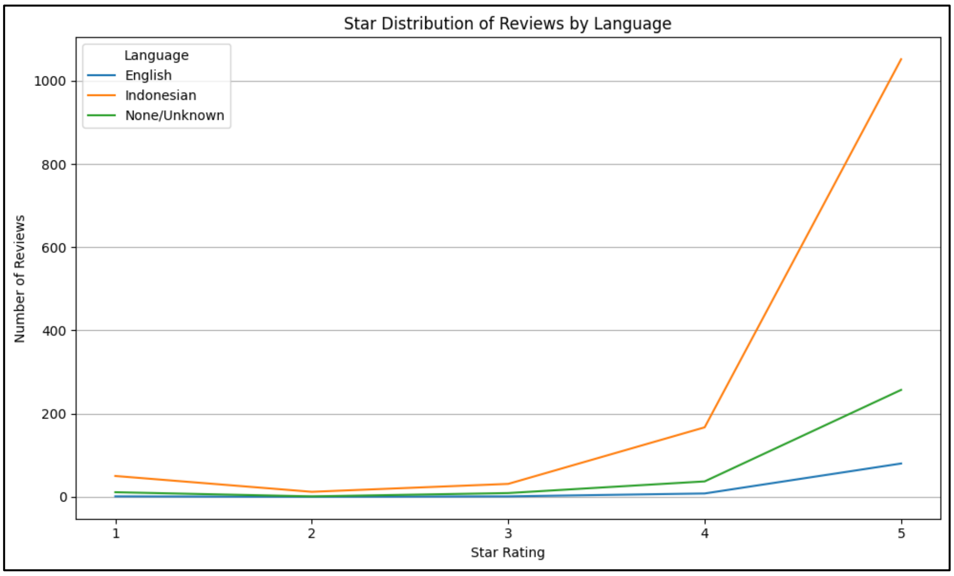

Figure : 6 Star Distribution of Reviews by Language.

4.3 Model Performance Testing

4.3.1 Model Selection

To evaluate sentiment classification performance, we tested three pre-trained multilingual models from Hugging Face, as identified in Chapter 2. These models were selected based on their popularity and active usage statistics as of October 2024. The chosen models are:

1. lxyuan/distilbert-base-multilingual-cased-sentiments-student (2.05M)
2. crypter70/IndoBERT-Sentiment-Analysis (9.49K)
3. carant-ai/xlm-roberta-sentiment-base (243)

The evaluation was conducted using our extreme sampling dataset, which consists of 1,338 reviews. The dataset includes only 1-star (negative) and 5-star (positive) reviews, ensuring clear sentiment boundaries for assessment.

4.3.2 Performance Testing

In this study, we curated a dataset of 5,297 Google Maps reviews from a neighboring university. During preprocessing, we addressed missing values by removing entries with **NaN (Not a Number)** values in the review text field. The dataset size before and after NaN removal is shown in Table 6.

Table 6 : NaN Removal Result

Dataset State	Total Reviews
Before NaN removal	5297
After NaN removal	1566

After cleaning the dataset, we applied an extreme review sampling technique commonly used in sentiment analysis research to establish clear sentiment boundaries. This approach involved filtering reviews at the opposite ends of the rating spectrum:

- Positive sentiment: Reviews with 5-star ratings
- Negative sentiment: Reviews with 1-star ratings

By using this method, we ensured that the dataset consisted only of strongly positive and strongly negative reviews, minimizing ambiguity in sentiment classification. The final dataset size after applying extreme sampling is presented in Table 7.

Table 7 : Extreme Sampling Result

Filtering Step	Total Reviews
After extreme sampling	1338

The sentiment distribution in the final dataset is displayed in Table 8

Table 8 : Sentiment Result

Sentiment	Review Count
Positive (5★)	1306
Negative (1★)	32

4.3.3 Model Evaluation

Each model was evaluated based on the following performance metrics:

- Accuracy: The percentage of correctly classified reviews.
- Precision: The proportion of correctly predicted positive reviews out of all predicted positive reviews.
- Recall: The proportion of correctly predicted positive reviews out of all actual positive reviews.
- F1 Score: The harmonic mean of precision and recall, balancing both metrics.

Table 9 : Performance Evaluation

Model	Accuracy	Precision	Recall	F1 Score
lxyuan/distilbert	0.9297	0.9715	0.9297	0.9477
crypter70/IndoBERT	0.8386	0.9793	0.8386	0.9010
carant-ai/xlm-roberta	0.8004	0.9793	0.8004	0.8777

Table 9 presents a comparative summary of the models' performance across these metrics. Among the evaluated models, the lxyuan/distilbert-base-multilingual-cased-sentiments-student model achieved the highest overall performance, with an accuracy of 92.97% and an F1 Score of 94.77%. Its high precision score of 97.15% indicates strong capability in minimizing false positives, while the recall of 92.97% shows that the model effectively captured most actual positive reviews. These results confirm that the lxyuan/distilbert model is the most reliable and well-balanced choice for multilingual sentiment classification in this study.

4.4 Data Processing

This chapter presents the results of the sentiment analysis performed using the chosen pre-trained model from the previous step. The findings are analyzed in terms of sentiment distribution, confidence score distribution, and the relationship between text length and model confidence. Furthermore, a discussion of the implications of these results is provided.

4.4.1 Sentiment Analysis

The sentiment analysis was conducted on a dataset consisting of 1,717 text entries, using the lxyuan/distilbert-base-multilingual-cased-sentiments-student model. The analysis was performed on a CPU, achieving an average processing speed of approximately 12.08 texts per second.

Table 10 : Sentiment Distribution

Sentiment	Sentiment Count
Positive	1511 (88.00%)
Negative	196 (11.42%)
Neutral	10 (0.58%)

As shown in Table 10, most reviews were classified as positive (88.00%) and negative (11.42%), while only 0.58% were neutral. This small neutral proportion may result from model bias toward polarized sentiments, especially with short or ambiguous texts. The researcher views this as a margin of error, given that the model was not fine-tuned for balanced three-class sentiment detection. Future work could address this by improving threshold calibration, increasing neutral training samples, or applying techniques like aspect-based sentiment analysis for better neutral classification.

4.4.2 Confidence Score Analysis

The confidence scores of the model's predictions were analyzed to assess the reliability of sentiment classification. The average confidence score across all predictions was 0.7285, indicating that, on average, the model assigned a relatively high confidence level when classifying reviews.

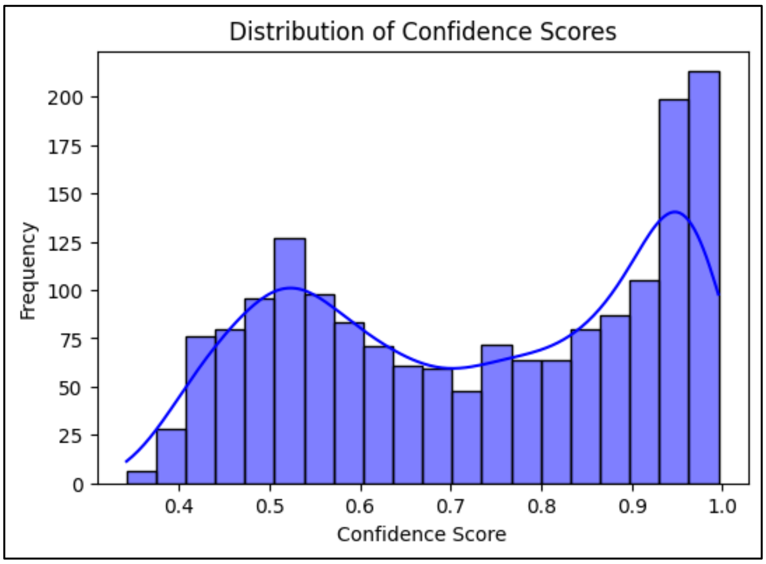

Figure : 7 Distribution of Confidence Scores.

Figure 7 illustrates the distribution of confidence scores across the entire dataset. As seen in the figure, the scores are skewed towards the higher end, with a large portion of predictions falling between 0.7 and 1.0. This indicates that the model often made predictions with strong certainty. Out of 1,717 total predictions, 937 reviews (54.57%) had a confidence score above 0.7, signifying high-confidence predictions. The remaining 780 predictions (45.43%) had scores below the 0.7 threshold, suggesting cases where the model was less certain. These lower-confidence predictions may require further analysis or represent more ambiguous sentiment content, such as sarcasm, neutral phrasing, or mixed language use.

4.4.3 Text Length and Confidence Relationship

The distribution of review text lengths was examined to assess variation in input data. The analysis revealed that most reviews are relatively short, with a large portion containing fewer than 25 words. However, a small number of reviews are significantly longer, indicating variability in how users express their opinions.

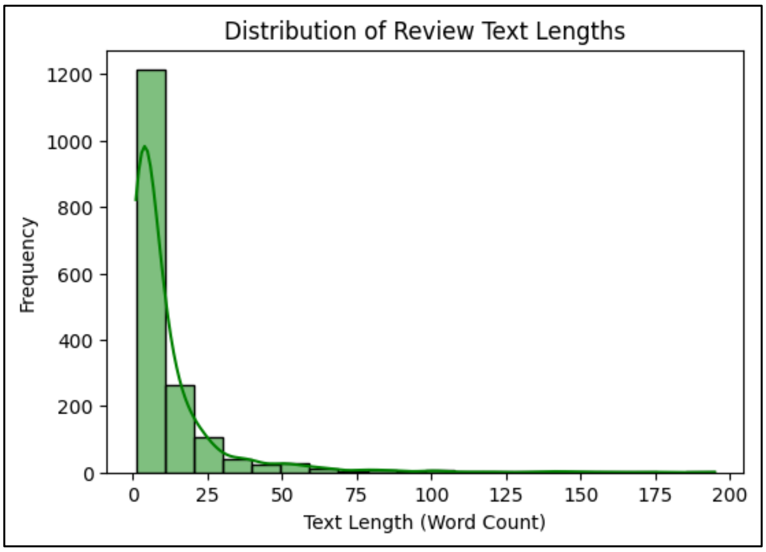

Figure : 8 Distribution of Review Text Lengths.

Figure 8 shows that the distribution of review lengths is right-skewed, with the majority being short texts. This suggests that the dataset largely consists of brief, user-generated comments, which may limit the amount of contextual information available for accurate sentiment classification. The dominance of short reviews poses challenges for transformer-based models that rely on sufficient context to make precise predictions. To explore the impact of review length on model performance, a correlation analysis was performed between text length and prediction confidence. The correlation coefficient obtained was 0.1253, indicating a weak positive relationship. While longer reviews tend to yield slightly higher confidence scores, the effect is minimal and does not significantly influence the overall model reliability.

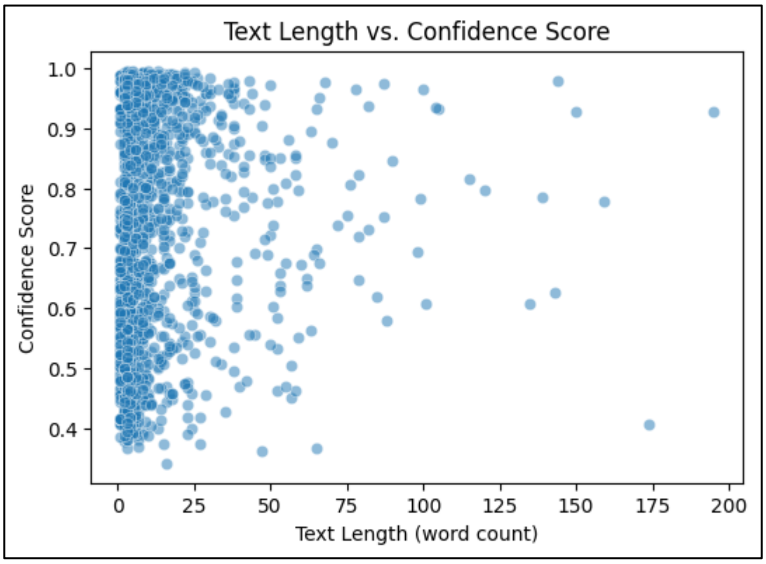

Figure : 9 Correlation between Text Length and Confidence.

As shown in Figure 9, there is a slight tendency for longer texts to be associated with higher confidence scores. However, the correlation is not strong enough to establish a definitive relationship. This suggests that while text length may contribute marginally to confidence, other factors likely play a more substantial role in the model's certainty.

4.4.4 Derived Feature Engineering

Beyond sentiment classification, additional feature engineering was conducted to uncover deeper insights from the dataset. These derived features include review quality indicators, temporal trends, and content-based patterns.

Table 11 : Review Quality Indicators

Review Quality Indicators	
Average Review Length	11.36 words
Average Reviews per User	1.15 reviews
Potential Spammers	7 users with more than 4 reviews

Table 11 presents several review quality indicators. On average, users wrote 11.36 words per review, suggesting that most feedback is brief and concise. Each user contributed approximately 1.15 reviews, with the majority submitting only once. Notably, 7 users were identified as potential spammers, as they posted more than four reviews, significantly above the average contribution.

The dataset exhibits relatively short review lengths on average, indicating that most users provide brief feedback. Additionally, the majority of users contribute only one review, while a small number of users exhibit unusually high review counts.

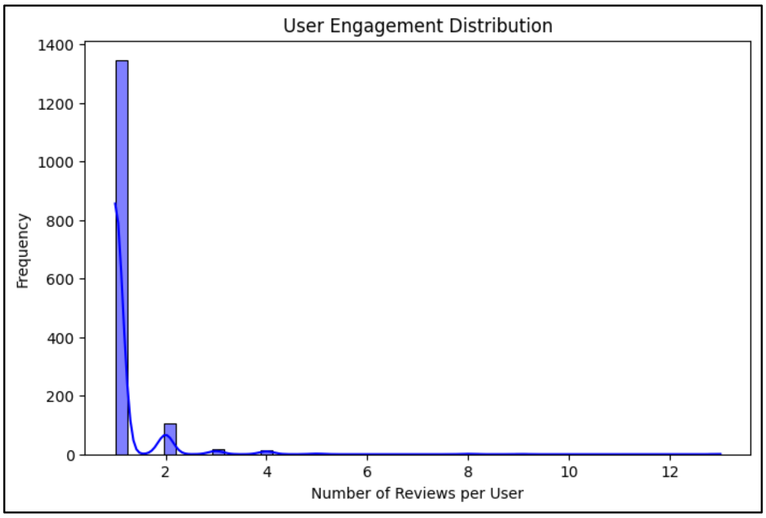

Figure : 10 User Engagement Distribution.

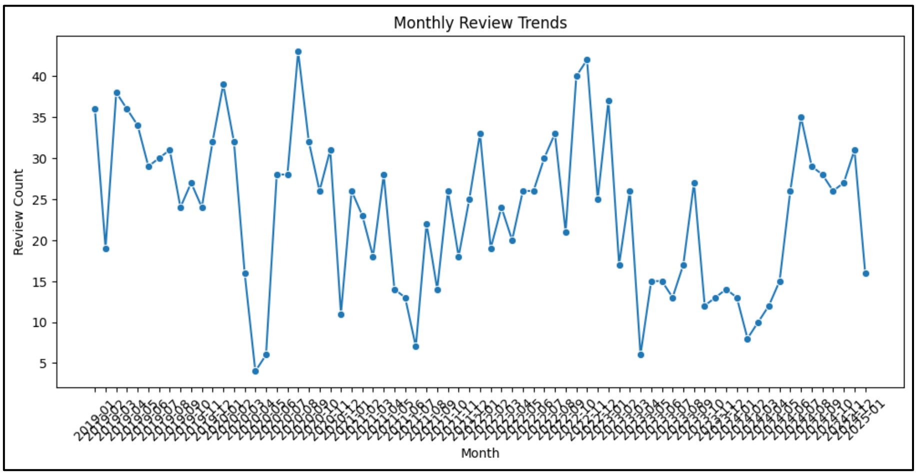

Figure : 11 Monthly Review Trends.

These insights suggest periodic fluctuations in user engagement, with certain months experiencing surges in review activity while others see minimal contributions.

Figure 10 illustrates the distribution of user engagement, further confirming that most contributors only post once, with a long-tail distribution of users who post more frequently. In addition, Figure 11 displays the monthly review trends, highlighting temporal patterns in submission volume. Based on the analysis:

- Months with >35 reviews: 8 months, with peaks in January 2019, March 2019, April 2019, January 2020, August 2020, October 2022, November 2022, and January 2023.
- Months with <10 reviews: 5 months, notably April 2020, May 2020, July 2021, April 2023, and February 2024.

These patterns suggest periodic fluctuations in user engagement, potentially influenced by external factors such as academic calendars or institutional events.

4.5 Interactive Data Visualization System

A comprehensive web-based visualization system was developed to present sentiment analysis results through an intuitive interface, utilizing modern web technologies and efficient data handling mechanisms

4.5.1 Dashboard Implementation

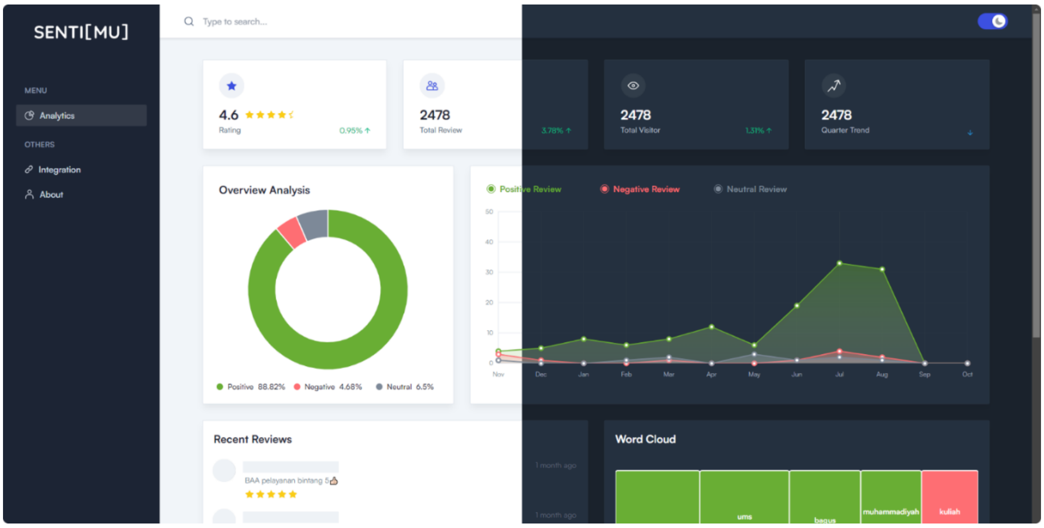

Figure : 12 Fully Developed Dashboard.

The interactive dashboard was developed using a modern tech stack comprising Next.js, TypeScript, and Tailwind CSS. This combination was selected to ensure high performance, maintainability, and scalability of the platform. Next.js, in particular, supports server-side rendering and static site generation, making it ideal for building responsive web applications that require real-time data integration and seamless future enhancements. The dashboard, titled SentiMu, integrates both key performance indicators (KPIs) and analytical visualizations to deliver actionable insights from sentiment analysis of Google Maps reviews. This dashboard is accessible to UMS stakeholders as a simple monitoring tool to track current public sentiment. There is no fixed schedule or required interaction; users can freely use the dashboard at any time weekly, biweekly, or monthly as needed. The information provided helps support evaluation and internal decision-making, particularly in understanding user satisfaction trends. The implemented features include:

1. Key Performance Indicators (KPIs)
    a. Aggregate star rating metrics

       Computational mean of cumulative star ratings assigned by reviewers to the University Muhammadiyah Surakarta (UMS).

    b. Total review counter

Quantitative measurement of total reviews, encompassing both textual and non-textual submissions.

c. Website visitor analytics

Quantitative analysis of SentiMu platform visitor traffic.

d. Current quarter sentiment trend analysis

Temporal analysis of predominant public sentiment, highlighting the modal sentiment category (positive, negative, or neutral).

2. Analytical Visualizations

a. Recent review feed

Dynamic display of five most recent UMS reviews, facilitating immediate feedback monitoring.

b. Sentiment distribution overview

Proportional representation of sentiment categorization across the UMS feedback spectrum.

c. 12-month rolling time series analysis

Rolling 12-month temporal distribution of aggregated sentiment metrics.

d. Interactive word cloud representation

Dynamic word cloud representation with keyword filtering capabilities.

e. Top reviewer identification panel

Identification of multiple-review contributors, serving dual functionality as a spam detection mechanism.

4.5.2 API Architecture

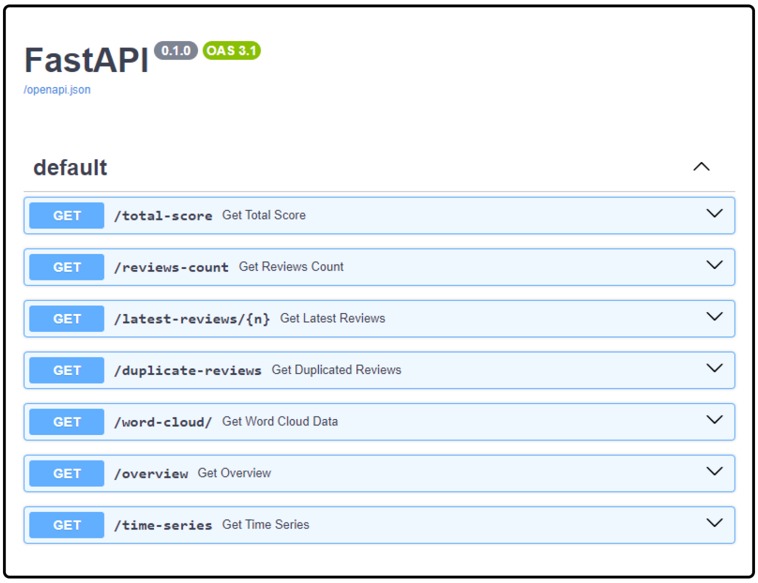

Figure : 13 Backend Endpoint Architecture.

As shown in Figure 13, the backend architecture is built on a FastAPI-based infrastructure designed to manage both data serving and processing tasks efficiently. The system includes multiple RESTful API routes that are responsible for delivering:

1. Endpoint Design
   a. Aggregated metrics retrieval
   b. Time series data endpoints
   c. Text search functionality
   d. Word frequency analysis(Word Cloud)
2. Data Processing
   a. Batch processing implementation
   b. CSV data transformation
   c. Sentiment analysis results serving

4.6 Summary of Finding

This study focused on sentiment analysis of Google Maps reviews related to Universitas Muhammadiyah Surakarta (UMS) using a pre-trained sentiment classification model. The research involved collecting, preprocessing, analyzing, and visualizing review data to extract meaningful insights about user sentiments.

Key findings from the study include:

1. Data Collection: A total of 4,665 Google Maps reviews were collected from 25 university-related locations. After preprocessing, 1,717 clean reviews remained for analysis.
2. Language Distribution: The majority of reviews (76.4%) were in Indonesian, with a smaller proportion in English (5.2%) and unknown/mixed-language reviews (18.4%).
3. Sentiment Distribution: The sentiment analysis revealed that 88.00% of reviews were classified as positive, 11.42% as negative, and only 0.58% as neutral.
4. Model Performance: Among three evaluated models, the lxyuan/distilbert-base-multilingual-cased-sentiments-student model demonstrated the highest accuracy (92.97%) and F1-score (94.77%), making it the most reliable choice for sentiment classification.
5. Confidence Score Analysis: The model showed an average confidence score of 0.7285, with 54.57% of predictions having a confidence score above 0.7, indicating reliable classification.
6. Text Length Analysis: The average review length was 11.36 words, with a weak positive correlation (0.1253) between text length and confidence score, suggesting longer texts slightly enhance classification confidence.
7. Temporal Trends: Review activity fluctuated over time, with peaks in certain months (e.g., January 2019, January 2020, and January 2023) and minimal activity in others (e.g., April 2020 and April 2023).

8. User Engagement Patterns: Most users submitted only one review, while seven users were identified as potential spammers (posting more than four reviews each).

4.7 Implications and Contributions

The findings of this study provide several key implications for both sentiment analysis research and institutional reputation management:

1. Academic Institutions: UMS can leverage sentiment analysis insights to identify strengths and weaknesses in campus facilities and student experiences.
2. Natural Language Processing (NLP) Research: This study demonstrates the effectiveness of pre-trained multilingual models in analyzing Indonesian-language user-generated content.
3. Automated Review Monitoring: Institutions can utilize sentiment classification models to track real-time sentiment trends and respond proactively to negative feedback.
4. This research also offers a novel contribution by implementing a transformer based sentiment analysis pipeline on a multilingual dataset of Google Maps reviews from a higher education context in Indonesia, without requiring model fine-tuning. The results are further integrated into a fully developed, real-time dashboard, enabling stakeholders to make timely and data-informed decisions.

4.8 Limitations

Despite the promising results, this study has some limitations:

1. Data Imbalance: The dataset was predominantly positive (88%), which may have impacted the model's ability to distinguish between negative and neutral sentiments.
2. Limited Feature Engineering: While some derived features (e.g., review length, temporal trends) were analyzed, additional linguistic and contextual features could further enhance sentiment classification.
3. Short Review Texts: The majority of reviews were brief, potentially limiting the model's ability to capture nuanced sentiment.

4.9 Future Work

Future research can expand upon this study in several ways:

1. Dataset Expansion: Collecting a larger and more balanced dataset, including more neutral and negative reviews, could improve sentiment classification.
2. Advanced Feature Engineering: Incorporating semantic and contextual features, such as emotion detection and aspect-based sentiment analysis, can refine sentiment insights.
3. Real-Time Sentiment Tracking: Implementing an automated monitoring system for continuous sentiment analysis can help universities respond to feedback dynamically.

4. Model Fine-Tuning: While this study utilized pre-trained models without fine-tuning, future research could explore customized model training on Indonesian-language datasets for improved accuracy.

This chapter has demonstrated the implementation and outcomes of sentiment analysis on Google Maps reviews of UMS. The final chapter consolidates these findings into a coherent conclusion, highlighting both theoretical and practical significance.

5. Conclusion

This study demonstrates the effectiveness of applying transformer-based multilingual sentiment analysis models to Google Maps reviews of Universitas Muhammadiyah Surakarta (UMS). By analyzing user-generated content, the research identifies key sentiment patterns, temporal trends, and behavioral insights that can inform institutional strategies and service improvements. The main contributions of this study include the integration of multilingual NLP models in the context of higher education, the development of a real-time dashboard for monitoring sentiment trends, and the extraction of user-driven quality indicators. These findings contribute to both academic discourse and practical applications, particularly in enhancing data-informed decision-making in university settings. Despite these contributions, the study has limitations, including reliance on publicly available reviews that may be subject to bias, and the absence of socio-demographic data that could provide deeper context to the sentiments expressed. Future research could expand on this work by incorporating data from multiple platforms, enriching datasets with user profiles and location metadata, and applying fine-tuning approaches to improve model performance in Bahasa Indonesia and similar low-resource languages.

Acknowledgments

The author would like to thank Universitas Muhammadiyah Surakarta, especially the Informatics Study Program, which has provided support during this research process. The author would also like to thank the supervisors, Mr. Endang Wahyu Pamungkas and Sohail Akhtar, who have provided guidance, direction, and input that are very helpful in the preparation of this report. The author is also grateful to all those who have contributed to data collection and system development.

Funding Information

The authors declare no funding was received for this study.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Ethical Approval

This study did not involve human or animal subjects.

Data Availability

Provide information on how data can be accessed:

- The dataset used is available at [https://github.com/SentiMu].
- https://maps.app.goo.gl/LTurYFKgNZTprQY97 (UMS)
- https://maps.app.goo.gl/yaQVGsF7KVJMuXRs9 (UMY)
- https://maps.app.goo.gl/EJEcSArUGK4PAJM98 (UNS)

References

[1] B. Tucker, "Disruptive Trends in Student Experience Evaluations and Implications for Academic Staff Wellbeing," *Assessing and Enhancing Student Experience in Higher Education*, pp. 261–286, Nov. 2021, doi: 10.1007/978-3-030-80889-1_11.

[2] N. Gronberg, A. Knutas, T. Hynninen, and M. Hujala, "Palaute: An online text mining tool for analyzing written student course feedback," *IEEE Access*, vol. 9, pp. 134518–134529, 2021, doi: 10.1109/ACCESS.2021.3116425.

[3] T. AYUB and S. Ganaie, "User Opinion About SPS Public Library: A Study Of Google Map Reviews," *Library Philosophy and Practice (e-journal)*, Jun. 2023, Accessed: Oct. 03, 2024. [Online]. Available: https://digitalcommons.unl.edu/libphilprac/7772

[4] Á. Borrego and M. Comalat Navarra, "What users say about public libraries: an analysis of Google Maps reviews," *Online Information Review*, vol. 45, no. 1, pp. 84–98, Jan. 2021, doi: 10.1108/OIR-09-2019-0291/FULL/XML.

[5] A. M. Khan and F. A. Loan, "Exploring the reviews of Google Maps to assess the user opinions about public libraries," *Library Management*, vol. 43, no. 8–9, pp. 601–615, Oct. 2022, doi: 10.1108/LM-05-2022-0053/FULL/XML.

[6] C. Yang *et al.*, "Online User Review Analysis for Product Evaluation and Improvement," *Journal of Theoretical and Applied Electronic Commerce Research 2021, Vol. 16, Pages 1598-1611*, vol. 16, no. 5, pp. 1598–1611, May 2021, doi: 10.3390/JTAER16050090.

[7] W. Widayat, "Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 3, pp. 1018–1026, Jul. 2021, doi: 10.30865/MIB.V5I3.3111.

[8] H. Imaduddin, F. Yusfida A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *IJACSA) International Journal of*

Advanced Computer Science and Applications, vol. 14, no. 8, p. 2023, 2023, Accessed: Oct. 22, 2024. [Online]. Available: www.ijacsa.thesai.org

[9] H. T. Atmoko, N. D. Wahyu Cahyani, and S. Kurniawan, "Grouping and Categorizing Data from Social Networking Applications for Forensic Analysis," *2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)*, pp. 67–72, Aug. 2024, doi: 10.1109/ICOABCD63526.2024.10704283.

[10] D. Chicco, L. Oneto, and E. Tavazzi, "Eleven quick tips for data cleaning and feature engineering," *PLoS Comput Biol*, vol. 18, no. 12, p. e1010718, Dec. 2022, doi: 10.1371/JOURNAL.PCBI.1010718.

[11] A. Aspin, "Data Cleansing," *Pro Data Mashup for Power BI*, pp. 293–320, 2022, doi: 10.1007/978-1-4842-8578-7_10.

[12] Y. Gaur *et al.*, "Streaming, Fast and Accurate on-Device Inverse Text Normalization for Automatic Speech Recognition," *2022 IEEE Spoken Language Technology Workshop, SLT 2022 - Proceedings*, pp. 237–244, 2023, doi: 10.1109/SLT54892.2023.10022543.

[13] C. Chen *et al.*, "WhatsNext: Guidance-enriched Exploratory Data Analysis with Interactive, Low-Code Notebooks," *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 209–214, Oct. 2023, doi: 10.1109/VL-HCC57772.2023.00033.

[14] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Caleanu, "Natural language processing with transformers: a review," *PeerJ Comput Sci*, vol. 10, p. e2222, Aug. 2024, doi: 10.7717/PEERJ-CS.2222/TABLE-3.

[15] S. Hazmoune and F. Bougamouza, "Using transformers for multimodal emotion recognition: Taxonomies and state of the art review," *Eng Appl Artif Intell*, vol. 133, p. 108339, Jul. 2024, doi: 10.1016/J.ENGAPPAI.2024.108339.

[16] W. Jiang *et al.*, "An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry," *Proceedings - International Conference on Software Engineering*, pp. 2463–2475, 2023, doi: 10.1109/ICSE48619.2023.00206.

[17] "Distilbert Base Multilingual Cased Sentiments Student · Models · Dataloop." Accessed: Jan. 17, 2025. [Online]. Available: https://dataloop.ai/library/model/lxyuan_distilbert-base-multilingual-cased-sentiments-student/

[18] S. Aras, M. Yusuf, R. Ruimassa, E. Agustinus, B. Wambrauw, and E. B. Palalangan, "Sentiment Analysis on Shopee Product Reviews Using IndoBERT," *Journal of Information Systems and Informatics*, vol. 6, no. 3, pp. 1616–1627, Sep. 2024, doi: 10.51519/JOURNALISI.V6I3.814.

[19] "carant-ai/xlm-roberta-sentiment-base · Hugging Face." Accessed: Jan. 17, 2025. [Online]. Available: https://huggingface.co/carant-ai/xlm-roberta-sentiment-base

[20] H. Imaduddin, Widyawan, and S. Fauziati, "Word embedding comparison for Indonesian language sentiment analysis," *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, pp. 426–430, Mar. 2019, doi: 10.1109/ICAIIT.2019.8834536.

[21] E. Blanco-Mallo, J. Carneiro, G. Marreiros, B. Remeseiro, and V. Bolon-Canedo, "When the best reviews are not placed between extremes," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2022-July, 2022, doi: 10.1109/IJCNN55064.2022.9892131.

[22] Y. Zhang, Y. Li, and J. Liu, "Unified Efficient Fine-Tuning Techniques for Open-Source Large Language Models," Jul. 2024, doi: 10.21203/RS.3.RS-4660140/V1.

[23] A. Rossouw and H. Smuts, "Key Principles Pertinent to User Experience Design for Conversational User Interfaces: A Conceptual Learning Model," *Lecture Notes in Computer*

Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 14099 LNCS, pp. 174–186, 2023, doi: 10.1007/978-3-031-40113-8_17.
[24] Z. Rucks-Ahidiana, "Content Analysis," *Doing Good Qualitative Research*, pp. 361–372, Apr. 2024, doi: 10.1093/OSO/9780197633137.003.0031.