DOI: 10.22146/jnteti.v14i4.22038

Sentiment Analysis Review Threads Google Play Store with RoBERTa Model

Natan Kharisma A[1], **Dewi Lestari**[1], **Gatot T Pranoto**[1]

[1] Informatics Engineering, Faculty of Engineering, Trilogi University, Jakarta 12760, Indonesia

[Received: 28 June 2025, Revised: 19 August 2025, Accepted: 24 September 2025]
Corresponding Author: Dewi Lestari (email: dewy24@trilogi.ac.id)

ABSTRACT — The rapid development of internet technology globally, including in Indonesia, has drastically changed communication and interaction patterns between individuals. One impact is seen in the increasing use of text-based social media applications, such as Threads, developed by Meta. Within a short time, Threads managed to attract millions of users. However, the large number of user reviews on the Google Play Store presents its own challenges, particularly in manual sentiment analysis, which is very time-consuming and prone to bias. This research aims to overcome these challenges by implementing a variant of bidirectional encoder representations from transformers (BERT), the robustly optimized BERT pretraining approach (RoBERTa) model, which has been optimized for natural language processing. The research process followed the cross-industry standard process for data mining (CRISP-DM) framework, including several main stages: understanding the business context, data exploration and model building preparation, performance evaluation, and model deployment. Data were obtained directly from the Google Play Store and then cleaned through deduplication, normalization, and tokenization stages. The RoBERTa model demonstrated strong performance, with an accuracy of 88%. Precision was recorded at 92% for positive sentiment and 81% for negative sentiment, while recall was at 88% and 87%, respectively. The F1 score was also high, at 90% for positive and 84% for negative sentiment. When compared to algorithms like naïve Bayes and support vector machine (SVM), RoBERTa proved superior. This research opens opportunities for exploring other transformer models or using ensembles to improve performance in the future.

KEYWORDS — Sentiment Analysis, Threads App, Google Play Store, BERT Method, User Reviews, RoBERTa Model.

I. INTRODUCTION

The number of internet users continues to rapidly rapid grow. According to the latest report from We Are Social, the number of global internet users is estimated to reach 5.56 billion in 2025 [1]. Meanwhile, the world population is estimated to reach 8.2 billion in early 2025. In Indonesia, internet users have reached 221 million, which is 79.5% of the total population. Indonesia is among the countries with the highest number of internet users in the world. According to information from UNICEF, every half a second, a child around the world explores the internet for the first time. In Indonesia, the total number of internet users has also reached 221 million, which represents 79.5% of the total population. Interestingly, 9.17% of these users are under the age of 12, which makes the younger generation more vulnerable to risks related to cyberspace [1].

The Threads app, a text-based social media platform developed by Meta, allows users to share photos, videos, and text messages with close friends. It is similar to Twitter (X) and has attracted a great deal of attention from people from various backgrounds. According to a statement by Meta CEO Mark Zuckerberg, the number of Threads users has exceeded 100 million in just five days [2]. This app has received various reviews from users. Many of the services and campaigns carried out have generated criticism and suggestions from users, expressed through comments on the Threads application review. There are many constructive or destructive comments and read and classify each comment is time-consuming and ineffective [3]. Data from the Google Play Store shows a 4.5 out of 5-star rating. Despites its high rating, users have reported various issues, such as bugs, crashes, spam, ads, and privacy. Hence, user reviews and ratings exhibit inconsistencies. This research focuses on analyzing user sentiment regarding the Threads app [4]. Despite its limitations, sentiment analysis remains a valuable tool for businesses and individuals to gather information and make decisions based on public opinion [5].

Sentiment analysis is an effective method for understanding user perceptions using natural language processing, text analysis, computational linguistics, and biometrics approaches to systematically identify, extract, measure, and analyze emotional states and subjective information [6]. In previous research, sentiment analysis on the Threads app using the naïve Bayes algorithm demonstrated an accuracy of 73%, with a precision value of 70% and a recall of 99%. The naïve Bayes method was chosen because of its processing speed and ability to handle simple and easy-to-implement data of 1,500 text data, but it has limitations in capturing more complex contexts in the text [7]. Previous research also used the support vector machine (SVM) algorithm as the primary method for classifying review sentiment into positive and negative. Sentiment analysis of Threads app reviews was conducted using 2,000 review samples. After removing duplicate reviews, 1,429 reviews remained for further analysis. Of these, 1,103 reviews were categorized as positive, while 326 reviews were categorized as negative. Evaluation of the sentiment analysis results using the SVM algorithm showed that this model successfully achieved an overall accuracy level of 88%. For negative sentiment classification, the model produced a precision of 80%, a recall of 60%, and an F1 score of 69%. On the other hand, in detecting positive sentiment, the model showed more optimal performance with a precision of 90%, a recall of 96%, and an F1 score of 93%. These results indicate that the SVM model is more effective in identifying reviews with positive sentiment compared to reviews with negative sentiment [8].

As an alternative, the bidirectional encoder representations from transformers (BERT) method developed by Google, offers a more sophisticated approach to sentiment analysis. More sophisticated based models such as BERT and robustly optimized BERT pretraining approach (RoBERTa) can better capture the relationships between words in a sentence, thus potentially significantly improving the accuracy of sentiment analysis. In addition, the RoBERTa architecture allows for effective fine-tuning, making it suitable for a wide range of applications, including text classification and anomaly detection in specialized domains [9], [10]. RoBERTa, as an advanced version of BERT, has been shown to capture the nuances and deeper meaning of user reviews, which often contain sarcasm or context that traditional methods like naïve Bayes and SVM cannot capture. So far, RoBERTa has delivered superior performance, especially in aspect-based sentiment analysis (ABSA) tasks, where the model can capture significant syntactic information, improving sentiment prediction accuracy. Various studies have demonstrated the performance of trained RoBERTa well with adaptations of previous state-of-the-art models across various datasets or languages, demonstrating its flexibility and efficiency on complex language understanding tasks [11].

The application of RoBERTa in sentiment analysis has shown promising results in various domains. Prior study analyzed public sentiment towards the development plan of the Indonesian Capital City (IKN) based on data from Twitter using the hybrid RoBERTa-gated recurrent unit (GRU) method [12]. This study found that negative sentiment dominated the conversation with 7,907 tweets, compared to 7,126 tweets with positive sentiment. This hybrid model achieved a test accuracy of 72.78%, demonstrating the effectiveness of this approach in classifying sentiment on social media data. According to [13], the study aimed to conduct sentiment analysis in classifying reviews of the Threads application on the Google Play Store using the BERT method. Researchers have previously conducted sentiment analysis on user reviews of the National Health Insurance (Jaminan Kesehatan Nasional, JKN) mobile application in the context of mobile application reviews, using the RoBERTa model. The model they developed demonstrated an accuracy of 98.68% on the tested dataset. A more in-depth analysis of 42,863 reviews revealed a relationship between ratings and sentiment, with positive reviews typically receiving high ratings, and negative reviews typically receiving low ratings. However, some inconsistencies were found, which can be attributed to factors such as user error in providing ratings and difficulties in natural language processing, such as sarcasm and mixed sentiment [14]. This research is expected to not only improve the accuracy of user review sentiment classification but also make a significant contribution to the development of sentiment analysis-based applications. Furthermore, the results are expected to enrich insights into user sentiment patterns towards mobile applications. This research will also add to the literature on sentiment analysis and the use of the BERT method in the context of mobile application reviews, as well as serve as a reference for further research in this area.

II. METHODOLOGY

In this study, the BERT algorithm was implemented using the cross-industry standard process for data mining (CRISP-DM) methodology. This method is widely used and consists of five main, interconnected stages: business understanding and data understanding, data preparation, modeling, evaluation, and deployment [15]. This approach was chosen because it provides a systematic structure for data analysis, particularly in classifying sentiments from user reviews of the Threads application. It consists of several stages that form the overall research flow. The first stage involved problem identification and problem formulation, followed by business and data understanding, which were necessary to gain insight into the characteristics and quality of the data in the study. Next, the data underwent the preparation stage, including data cleaning, transformation, and integration to ensure they were ready for use in modeling. Once the data were prepared, the modeling stage was conducted and subsequently evaluated. The final stage was deployment, where the model was implemented in a production environment or used to support decision-making.

A. IDENTIFICATION OF PROBLEMS

The Threads app, developed by Meta, has attracted significant attention, surpassing 100 million users in just five days since its launch. This rapid growth demonstrates the high level of user enthusiasm. However, the large number of incoming reviews poses a new challenge, as manually reading and classifying each comment is both inefficient and time-consuming. These comments can include both constructive criticism and negative feedback, which are important for further analysis. Therefore, an efficient and effective solution is needed to automate the sentiment analysis and comment classification process, so that important user information can be optimally utilized to improve the quality and user experience of the Threads app.

B. BUSINESS & DATA UNDERSTANDING

In the initial stage, there are two main aspects that need to be considered: business and data understanding. Business understanding focuses on identifying the main objectives of the research conducted on the Threads app. In addition, data understanding also includes important aspects such as identifying data sources and the amount of data to be used in the model. In the context of sentiment analysis research on reviews or comments on an application, understanding the origin of the data is crucial to ensure its validity and relevance to the research objectives. Data sources can come from various platforms, such as social media, discussion forums, or app stores, each of which has different characteristics in conveying Threads user opinions. In addition, researchers determined the amount of data needed to build an accurate and representative sentiment analysis model. In this study, the BERT method was applied to analyze sentiment from Threads user comments, thus also including an understanding of the characteristics and data formats suitable for the RoBERTa model input.

1) BUSINESS UNDERSTANDING

Threads app was launched in response to changes at Twitter (X), Meta, as a major player in the technology industry. Meta aims to capture the opportunities arising from shifting user preferences and the need for a platform that focuses more on intimate conversational interactions. Threads app is designed to provide a different social media experience, with an emphasis on more structured and focused text-based conversations. In this research, the focus is on analyzing sentiments regarding user reviews of the Threads app obtained from the Google Play Store platform. This sentiment analysis has significant business implications. By understanding user sentiment, application developers and stakeholders can gain valuable insights into Threads, strengths and weaknesses, identify areas for improvement, and formulate strategies to increase user satisfaction. The primary objective of this research is to classify user comments into three categories: positive, negative, and neutral. This classification allows for identifying overall sentiment trends, as well as identifying specific comments that reflect user issues or concerns. Furthermore, sentiment analysis can also help in identifying the features or aspects of the application that users like or dislike the most, which can form the basis for developing new features or improving existing ones.

2) DATA UNDERSTANDING

The data understanding stage plays a crucial role in the success of this sentiment analysis research. At this stage, researchers conducted an in-depth exploration and comprehensive analysis of Threads app user review data collected from the Google Play Store platform. A good understanding of the data characteristics is crucial for the sentiment analysis model to run accurately and reliably. This study utilized user review text data as the primary source, which often contains complex language, slang, and diverse emotional expressions. Therefore, it is essential for researchers to understand the structure and variation of the language used in the reviews and identify any underlying patterns or trends. Furthermore, the data understanding stage also includes an analysis of data quality. In the context of sentiment analysis, inconsistent data contain missing values (NaN values) or has anomalies that can significantly impact the performance of machine learning models. Therefore, careful data preprocessing was carefully conducted, including data cleaning, handling missing values, and text normalization. Effective data preprocessing ensures that the data used to train the sentiment analysis model is clean, consistent, and representative.

Several crucial aspects that must be considered in understanding the data are data source and data amount. The data collected in this study were obtained from scraping from the Google Play Store platform. This platform has become a primary choice for Android smartphone users due to its reliability in providing a variety of applications accessible to many users worldwide [15]. The data to be scraped amounted to 2,000 data points in the “Most Relevant” category, including a rating of 1 to 5 on the Google Play Store platform. Researchers chose a relatively small amount of data to ensure the data compilation process was more efficient, considering that during the data cleaning stage, a lot of information would need to be removed. Some data points were deemed irrelevant and could not be included in the model [16].

C. DATA PREPARATION

Before data were used in the modeling process, a preparation phase was required to ensure its quality and relevance. This stage confirms the data can be processed correctly [17] and ensure accurate processing as well as adaptability to the specific requirements of the algorithm implemented in the model. Properly performed preprocessing stage (data preparation) improves data quality and structure, which directly impacts the performance and accuracy of the sentiment analysis model being developed. Data preparation consists of a series of interconnected and mutually supportive steps [18]. The process begins with data scraping process, data labeling process, case folding process, data cleaning process, tokenization process, stemming process, data exploration, splitting process, term frequency-inverse document frequency (TF-IDF) process, and smoothing process. The following is an in-depth explanation of each stage in the data preparation process.

1) DATA SCRAPPING STAGES

The initial stage of data preparation in this study involved a data collection process known as web scraping. Web scraping is an automated method used to systematically extract information from web pages or applications. In this context, it was applied to collect user reviews of the Threads app available on the Google Play Store. To facilitate this process, a Python library called google_play_scraper was used, providing functions that facilitate the extraction of review data, ratings, and other relevant information from the Google Play Store.

2) LABELING STAGE

The labeling process is crucial because machine learning models, particularly sentiment classification models, require labeled data to learn to distinguish between different sentiment categories. In this context, the model received user review text as input and produced a corresponding sentiment label as output. Therefore, the textual data must be converted into a numeric format that the model can understand, and sentiment labeling is the first step in this transformation. In data processing for machine learning-based models, particularly in sentiment analysis tasks, the input is generally text. However, because the model works with numeric data, the textual data must first be converted into numeric form so that it can be understood by the system. One method used to convert text into numbers is to label the data based on specific sentiment categories. In this labeling context, reviews or data with ratings of 1 to 2 were classified as negative sentiment, ratings of 3 were considered neutral, while ratings of 4 to 5 fell into the positive sentiment category.

3) CASE FOLDING

Case Folding is one of the important processes in text preprocessing stage, aiming to convert all characters in the text to lowercase [19]. This is done to ensure that words that are similar but have different capitalizations are considered the same word by the model. For example, the words “Application,” “APPLICATION,” and “application” are considered the same word after case folding.

4) CLEANING STAGE

Data Cleaning is process that ensures that the review text used for model training is high quality and free of irrelevant elements. With clean data, the model can learn sentiment patterns more accurately and produce more valid predictions.

5) TOKENIZATION PROCESS

Tokenization is a step to break down an input string into separate words or what are known as tokens. In general, the tokenization process divides a set of characters in a text into word units, by separating certain words that may or may not function as separators [20].

6) STEMMING

Stemming will filter words with the aim of converting words with affixes into their base form. For example, the word “memukul” will become “pukul.” At this stage, the “Sastrawi” library is used for stemming process [21]. This normalization is important because words with different affixes often have the same basic meaning, and stemming helps the model recognize these words as the same entity. The Sastrawi library, available in the Python programming language, was used to perform the stemming process. Sastrawi is specifically designed to handle stemming in Indonesian, applying relevant morphological rules.

7) EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a data analysis stage that aims to identify patterns, anomalies, test hypotheses, and validate assumptions before proceeding to the modeling stage [3]. This process is very important in ensuring that the data used is of adequate quality and meets the needs of further analysis.

8) TRAINING AND TEST DATA SPLITTING PROCESS

The process of splitting technique separates the data into a 70:30 ratio, 70% for training data and 30% for testing data. Data separation in the training process is crucial as it determines which data are used to train the model and which are used to test its performance. This separation allows the model to develop patterns based on the training data. The model will then be tested using the testing data to assess its ability to generalize to new data. One commonly used approach to split the data using a 70:30 ratio, which helps prevent "overfitting," a condition where the model is overly optimistic and almost reaches 100% in one epoch with a low loss rate [15]. It indicates that the model is overly optimistic about patterns in the training data and fails to generalize well. With proper separation, the model can learn more effectively and accurately when handling previously unseen data.

9) TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

In the data transformation stage, the Term Frequency-Inverse Document Frequency (TF-IDF) method is used to convert raw text that cannot be directly understood by computers into vector representations. In addition, TF-IDF also plays a role in assessing the level of importance of a word in a document based on its frequency of occurrence[22].TF-IDF works by giving lower weight to words with high frequency of occurrence, while words that appear less frequently will receive a higher weight. Frequently occurring words are also called stopwords, which are often ignored in text analysis to improve the efficiency and accuracy of data processing [23].

10) SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

Synthetic minority over-sampling technique (SMOTE) is a method used to address class imbalance issues in a dataset, particularly in sentiment analysis when there is a significant difference between the number of positive and negative reviews. This imbalance can cause the model to be more accurate in recognizing the majority class, while underrepresenting the minority class. Therefore, SMOTE is applied to increase the number of samples in the minority class so that the data distribution becomes more proportional. This process begins with dividing the processed dataset into two main parts: training data and test data. After this division, SMOTE is applied to the training data to increase the proportion of the minority class, thus creating a more balanced data distribution [24]. The result is training data increased by the presence of synthetic data generated through the oversampling process with SMOTE. Rather than simply duplicating existing samples, it creates new samples by interpolating between adjacent samples in the minority class. This method is effective in balancing the class distribution without causing overfitting [25].

D. MODELING

At this stage, the process entered the modeling phase, where data went through a series of preparation stages (such as cleaning, labeling, and transformation) was fed into a machine learning model for further analysis. The model used in this study was the development of BERT algorithm, namely RoBERTa, previously introduced by Google. BERT has a variant model that has gone through a training process using a large-scale dataset consisting of various types of text corpora, especially for the Indonesian language. Next, the optimizer setup was carried out. In this study, the optimizer used was Adam, allowing for determining the most optimal learning rate during the iteration stage of model training. The Adam optimizer was used in this study due to its ability to adapt the learning rate, with a value of 2e-5 applied to ensure optimal and efficient model parameter updates [26]. This model parameters can help in achieving faster convergence and more stable training [27]. Based on previous studies, this model could produce higher accuracy and providing more reliable predictions in the tested tasks. In this study, data with a neutral label (rating 3) was not included in the model training and testing process. This decision was made because the amount of available neutral data were relatively small and less representative, which could affect the class distribution unequally. Furthermore, the main focus of the study is the binary classification between positive and negative sentiment to optimize the accuracy and efficiency of the RoBERTa model. However, the presence of neutral data was still identified during the exploration stage to provide a comprehensive picture of the distribution of sentiment in user reviews.

E. EVALUATION

This stage focuses on the computational process in machine learning, which includes model implementation, data processing, algorithm selection, and the use of various predefined functions. In this study, the approach used involved scikit-learn, an open-source library developed within the Python programming language ecosystem. Scikit-learn was chosen because it provides various efficient tools and methods for building and evaluating machine learning models, thereby simplifying the process of data analysis and systematic algorithm implementation. This stage aims to establish evaluation metrics used as benchmarks in assessing the results of computational calculations in sentiment analysis [15]. Some of the main aspects analyzed in this stage are accuracy, precision, recall, and F1 score.

Accuracy is one of the easiest evaluation metrics used to assess a model's performance. Accuracy can be calculated using (1).

$$accuracy = \frac{TN+TP}{TP+FN+TN+FP}. \tag{1}$$

Precision is a metric that measures the accuracy of a model's positive predictions by comparing the number of predictions correctly classified as positive to the total number of predictions categorized as positive by the model. The formula for precision is presented in (2).

$$precision = \frac{TP}{TP+FP}. \tag{2}$$

Recall is an evaluation metric used to measure a model's ability to identify all data that should be classified as positive. The recall formula is presented in (3).

$$recall = \frac{TP}{TP+FN}. \tag{3}$$

The F1 score is a measure of model performance obtained from the harmonic mean of precision and recall. The formula for the F1 score is in (4).

$$f1-score = 2.\frac{precision.recall}{precision+recall}. \qquad (4)$$

F. DEPLOYMENT

The deployment phase in this research refers to the final process after the sentiment analysis parameters have been obtained. At this stage, the results were systematically documented in the form of a final project report. This final project report outlined how sentiment analysis was applied in real-world settings to benefit users or businesses. This implementation is expected to provide tangible benefits to users or businesses by assisting in data-driven decision-making [15].

III. RESULTS AND DISCUSSION

A. DATA UNDERSTANDING

This study utilized secondary data sources obtained directly from the internet using the Python programming language within the Google Colaboratory environment. The data collection process began with accessing the official Google Play Store website using the keyword “com.instagram.barcelona.” The data collected focused on reviews originating from users in Indonesia with the language designated as “ID” and only includes reviews categorized as “Most Relevant.” In this stage, a total of 2,000 records were collected [8]. In this study, data selection was carried out to train and test only negative and positive sentiments, while neutral sentiment data of 1,730 were not included in the model. The data, containing both positive and negative samples, were divided using a 70:30 split, with 70% of the data allocated for training and 30% for testing.

B. DATA PREPARATION

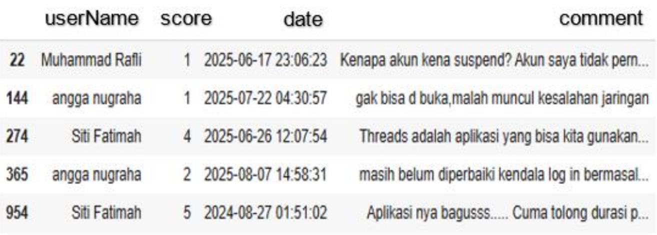
After data scraping, duplicate reviews were checked in the username column. From 2,000 reviews, rows with blank and duplicate values were removed in the username column, leaving 1,990 reviews to optimize the data for this study. The results of the data found and the results of removing duplicated data are shown in Figure 1.

The next step was labeling the review scores on a 1–5 rating scale. Ratings 1 to 2 were categorized as negative sentiment, 3 as neutral, and 4 to 5 as positive sentiment. This labeling process can be seen in Figure 2.

Subsequently, the cleaning process was initiated. The goal is to remove unnecessary attributes, irrelevant characters, and correct spelling errors. Therefore, the final data were cleaner, more structured, and ready for further processing. After this process, the modified data results are shown in Table I.

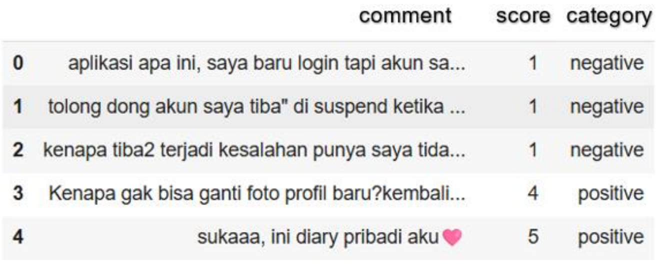
In the tokenization process, a sentence was broken down into parts by removing nonletter characters. This step aims to divide the text into smaller parts so that it can be further processed in natural language analysis. Tokenization facilitates the subsequent stemming stage to be more effective because the words have been separated into a more structured form. Table II shows the results of the tokenization process [8].

In the text preprocessing stage, stemming is a crucial step. This process aims to convert words with affixes or those that have undergone morphological changes to their base form, thereby improving the consistency of text data in further analysis. This aims to simplify word variations and reduce the feature dimensionality in text analysis. In this study, stemming was performed using the Sastrawi library available in the Python programming language. Sastra is a library specifically developed to implement the stemming process in the Indonesian language context. By applying relevant morphological rules, as shown in Table II, the information presents results relevant to the process described previously.



	userName	score	date	comment
22	Muhammad Rafli	1	2025-06-17 23:06:23	Kenapa akun kena suspend? Akun saya tidak pern...
144	angga nugraha	1	2025-07-22 04:30:57	gak bisa d buka,malah muncul kesalahan jaringan
274	Siti Fatimah	4	2025-06-26 12:07:54	Threads adalah aplikasi yang bisa kita gunakan...
365	angga nugraha	2	2025-08-07 14:58:31	masih belum diperbaiki kendala log in bermasal...
954	Siti Fatimah	5	2024-08-27 01:51:02	Aplikasi nya bagusss..... Cuma tolong durasi p...

Figure 1. Duplicated username data and selection data results.



	comment	score	category
0	aplikasi apa ini, saya baru login tapi akun sa...	1	negative
1	tolong dong akun saya tiba" di suspend ketika ...	1	negative
2	kenapa tiba2 terjadi kesalahan punya saya tida...	1	negative
3	Kenapa gak bisa ganti foto profil baru?kembali...	4	positive
4	sukaaa, ini diary pribadi aku 💗	5	positive

Figure 2. Labeling results based on score.

TABLE I
CLEANING AND SENTIMENT CLASSIFICATION

Before	After	Category
What application is this? I just logged in but my account has been suspended even though I just downloaded it.	What application is this? I just logged in but my account has been suspended even though I just downloaded it.	Negative
Why can’t I change my profile photo? Go back to the previous profile photo.	Why can’t I change my new profile photo back to my previous profile photo?	Positive
Why is it that every time I change my profile photo, it always goes back to the previous profile photo?	Why does every time I change my profile photo I always go back to the previous profile photo?	Neutral

The next step was to evaluate the data using an EDA approach, resulting in a chart diagram showing 1,076 positive data, 654 negative data, and 260 neutral data. In the context of this sentiment analysis, EDA is used to understand the distribution of sentiment in the user review dataset. The results of this analysis are visualized by showing the number of reviews for each sentiment category: positive, negative, and neutral. Figure 3 shows the bar chart and pie chart below.

C. MODELING

At this stage, the data went through a series of preprocessing steps. To produce the training procedure, the training and testing data were divided into the hugging face dataset format. Next, the model building process involved using a pretrained RoBERTa model. This approach leveraged the knowledge the model had learned on a large text corpus, allowing the model to adapt more quickly to specific classification tasks. The preprocessed dataset was then fed into the RoBERTa model to undergo a series of training and testing iterations. The results of this initial training process, including key metrics, are shown in Figure 4. These metrics provide insight into the progress and effectiveness of the model training [13].

global_step denotes the total number of training steps (batches passed) during training in this case, 228 steps. Training_loss is

TABLE II
TOKENIZING AND STEMMING

Before	Tokenizing	Stemming
What application is this? I just logged in but my account has been suspended even though I just downloaded it.	[application, login, account, suspend, download]	account login application suspend download
Why can't I change my profile photo? Go back to the previous profile photo.	[no, change, photo, profile, photo, profile, nya]	don't change the profile photo profile photo
Why is it that every time I change my profile photo, it always goes back to the previous profile photo?	[change, photo, profile, photo, profile, yg]	change profile photo profile photo that

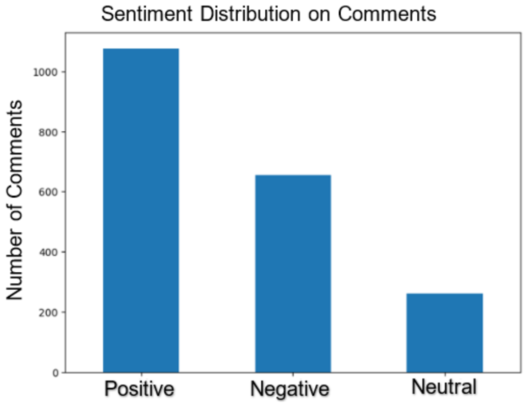
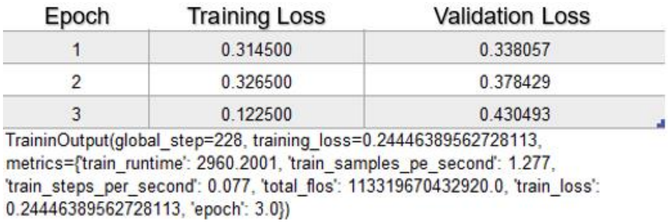


Figure 3. EDA dataset results.



Epoch	Training Loss	Validation Loss
1	0.314500	0.338057
2	0.326500	0.378429
3	0.122500	0.430493

```
TraininOutput(global_step=228, training_loss=0.24446389562728113,
metrics={'train_runtime': 2960.2001, 'train_samples_pe_second': 1.277,
'train_steps_per_second': 0.077, 'total_flos': 113319670432920.0, 'train_loss':
0.24446389562728113, 'epoch': 3.0})
```

Figure 4. Modeling results.

the average overall training loss (averaged across all epochs), in this case was 0.244. train_runtime is the total training time, which was 2,960 s (~49 min). train_samples_per_second is training speed in number of samples processed per second (1.227 samples/s). train_steps_per_second is training speed in batch steps per second (0.077 steps/s). The conclusion is that the model performance was less stable, because there was overfitting (validation loss increases at the end of the epoch).

D. EVALUATION

After going through an intensive training process, the sentiment classification model underwent a rigorous evaluation process to measure and assess its ability to predict previously unseen data. This evaluation aims to provide a more accurate picture of the model's performance when applied to real-world situations. The evaluation results were visualized using a confusion matrix, providing a graphical representation of the model's performance. This matrix shows the number of correct and incorrect predictions for each class, as shown in Figure 5.

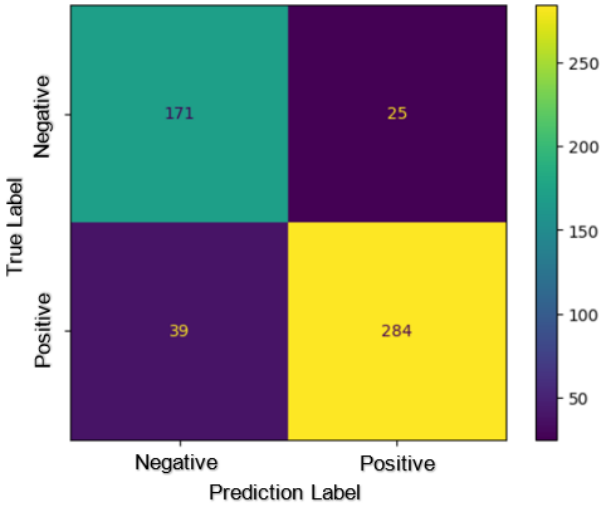
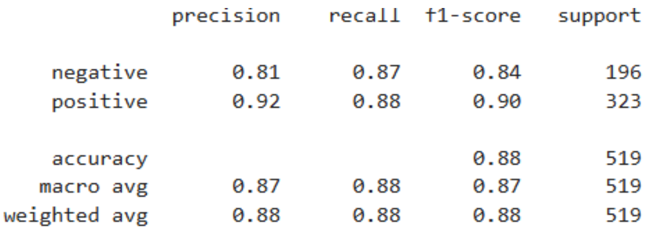


Figure 5. Confusion matrix.



```
              precision    recall  f1-score   support

    negative       0.81      0.87      0.84       196
    positive       0.92      0.88      0.90       323

    accuracy                           0.88       519
   macro avg       0.87      0.88      0.87       519
weighted avg       0.88      0.88      0.88       519
```

Figure 6. Evaluation results.

Figure 5 shows the confusion matrix resulting from the RoBERTa model evaluation on the test data. This matrix consists of four main components: 171 true negatives (TN), 25 false positives (FP), 39 false negatives (FN), and 284 true positives (TP). The horizontal axis represents the predicted labels, while the vertical axis shows the true labels. The colors in the matrix indicate the intensity of the data, with a gradient from dark purple (low values) to bright yellow (high values). This visualization provides a clear picture of the distribution of model predictions and the level of classification accuracy for each sentiment class. From the confusion matrix, a comprehensive evaluation report was obtained, covering various performance metrics in Figure 6.

Based on the evaluation results using a confusion matrix, this model achieved an accuracy rate of 88%. In addition, the precision was recorded at 92%, recall at 88%, and F1 score reached 90% for the positive category from a total of 323 data points analyzed. For negative results, the precision value was 81%, recall at 87%, and F1 score at 84%, based on a total of 196 data points. In addition, the macro average of 87% and weighted average of 88% indicated a balance in model performance across both classes. These results indicate that the BERT method provides slight advantage over naïve Bayes and SVM in classifying sentiment from user reviews of the Threads application. With higher accuracy, precision, recall, and F1 score values, the BERT method is proven to be superior in conducting sentiment analysis on this application review data, confirming its effectiveness in more accurate sentiment classification.

Comparisons with the naïve Bayes algorithm and SVM were derived from previous studies conducted on similar data but not retested on this research dataset. It aims to provide an overview of the RoBERTa model's performance from a literature perspective. However, differences in preprocessing approaches and data volumes across previous studies are important to note to avoid misinterpreting the results as a uniform empirical evaluation. The following model comparison is described in Table III.

The results come from previous studies and not from direct experiments in this study [7], [23].

TABLE III
MODEL COMPARISON RESULTS

Model	Sentiment	Accuracy	Precision	Recall	F1-Score
RoBERTa	Positive	88%	92%	88%	90%
	Negative		81%	87%	84%
NB	Positive	73%	70%	99%	82%
	Negative		94%	27%	42%
SVM	Positive	88%	83%	77%	80%
	Negative		79%	84%	81%

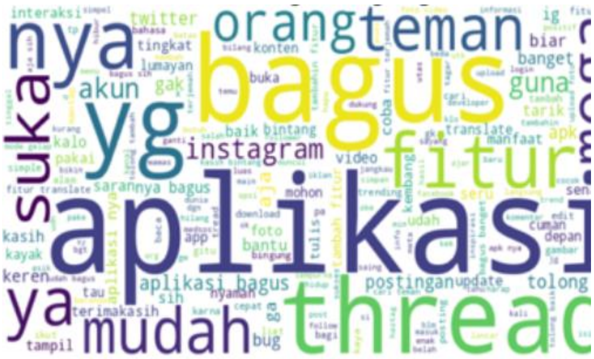


Figure 7. Wordcloud visualization and positive word frequency.

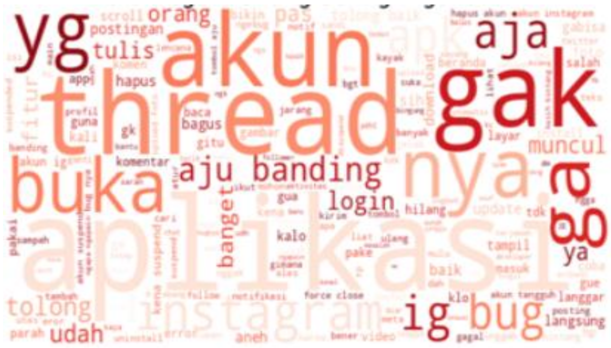


Figure 8. Wordcloud visualization and negative word frequency.

Next, visualizing words from each category using a wordcloud library approach produced an image where the size of each word indicated its frequency of occurrence in the text; words with higher frequency of occurrence are displayed in a larger size to highlight the term's dominance in the dataset. This visual data processing aims to provide a deeper and more intuitive understanding of the dominant themes and sentiments in user reviews. The visual data processing format is presented as follows.

1) POSITIVE WORDCLOUD

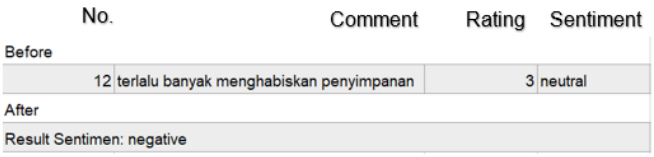
A positive sentiment wordcloud was generated from a collection of user reviews classified as positive regarding the Threads app. In this visualization, several words appear prominently, indicating their significant frequency of occurrence. Words such as "app," "good," and "threads" are noticeably dominant, indicating that users frequently use these words in positive contexts. These words reflect users, appreciation of the app's quality and features. This wordcloud visualization provides a glimpse into the aspects of the app that users value most. The visualization and frequency are shown in Figure 7.

2) NEGATIVE WORDCLOUDS

A negative sentiment wordcloud was generated from a collection of user reviews classified as negative regarding the Threads app. In this visualization, some words appear larger, indicating their significant frequency in negative contexts. Words such as "app," "account," and "appeal" are noticeably dominant, indicating that users frequently use these words when expressing dissatisfaction or criticism of the app, as visualized and shown in Figure 8.

E. DEPLOYMENT

Based on the modeling results performed using the RoBERTa model, this model can be applied in the decision-making process to classify user reviews. Thus, this model helps ensure consistency in review assessment, reducing the possibility of differences in interpretation between reviewers. This aspect can be beneficial for companies in designing future application development strategies. Figure 9 shows a comparison of the results before and after the model training process.



No.	Comment	Rating	Sentiment
Before			
12	terlalu banyak menghabiskan penyimpanan	3	neutral
After			
Result Sentimen: negative			

Figure 9. Before and after pretrained model.

IV. CONCLUSION

This study has successfully identified and classified user sentiment towards the Threads application using the BERT method. The classification report for the testing data using the RoBERTa model with a 70:30 data split demonstrated excellent performance in predicting the "negative" and "positive" classes with an overall accuracy of 88%. High precision and recall, as well as a balanced F1 score, indicate the model's ability to perform accurate classification. For further research, it is recommended to explore other methods such as the use of newer transformer models such as XLNet or ELECTRA or combine RoBERTa with ensemble learning techniques to improve performance. In addition, it is necessary to consider conducting further analysis of classification errors to identify patterns or data characteristics that may affect model performance, as well as exploring data augmentation techniques to address potential class imbalances. The analysis results show that the BERT method can provide higher accuracy in sentiment classification compared to traditional algorithms such as naïve Bayes and SVM.

CONFLICTS OF INTEREST

The authors have no personal interests related to the issues or topics discussed in this manuscript. The data collection procedures, methodology, and interpretation of the results were conducted without influencing any system. In other words, the validity and independence of the results presented are proven.

AUTHORS' CONTRIBUTIONS

Conceptualization, Natan Kharisma A. and Dewi Lestari; methodology, Natan Kharisma A.; software, Natan Kharisma

A; validation, Natan Kharisma A., Dewi Lestari, and Gatot T. Pranoto; formal analysis, Natan Kharisma A.; investigation, Natan Kharisma A.; resources, Natan Kharisma A.; data curation, Natan Kharisma A.; writing—original draft preparation, Natan Kharisma A.; writing—review and editing, Natan Kharisma A.; visualization, Natan Kharisma A; supervision, Dewi Lestari, and Gatot T. Pranoto; project administration, Natan Kharisma A.; funding acquisition, Dewi Lestari.

REFERENCES

[1] K. Wisnubroto, "Government,s commitment to protecting children in the digital space," indonesia.go.id. Accessed: Mar. 29, 2025. [Online]. Available: https://indonesia.go.id/kategori/editorial/9037/komitmen-pemerintah-melindungi-anak-di-ruang-digital?lang=1

[2] W. Meliani and D. Gustian, "Public opinion sentiment analysis of the Threads application on Twitter using the Naïve Bayes method," in Proc. Nat. Seminar Inf. Syst. Informatics Manage., Nusa Putra Univ., Jan. 2024, pp. 197–202. Accessed: Feb. 12, 2025. [Online]. Available: https://sismatik.nusaputra.ac.id/index.php/sismatik/article/view/260

[3] M.N. Akbar and N. Samrin, "Sentiment analysis of user comments on the Threads application on Google Playstore using the multinominal Naive Bayes classifier algorithm," Jagti, vol. 3, no. 2, pp. 21–29, Aug. 2023, doi: 10.24252/jagti.v3i2.67.

[4] M.F. Hanif, S.H. Wijoyo, and W.H.N. Putra, "Sentiment classification of Threads application reviews based on the Naive Bytes algorithm and root cause analysis method," J-PTIIK, vol. 8, no. 6, Jul. 2024. Accessed: Mar. 29, 2025. [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/13786

[5] M. Wankhade, A.C.S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[6] D. Naik, H. Sultana, and K.K. Jitendra, "Insight into sentimental analysis," J. Emerg. Technol. Innov. Res., vol. 7, no. 6, pp. 1561–1566, 2020. Accessed: Mar. 25, 2025. [Online]. Available: https://www.academia.edu/download/81006839/JETIR2006559.pdf

[7] N. Nurzaman, N. Suarna, and W. Prihartono, "Sentiment analysis of Threads app reviews on Google Playstore using the Naïve Bayes algorithm," Inf. Eng. Student J., vol. 8, no. 1, pp. 967–974, 2024, doi: 10.36040/jati.v8i1.8708.

[8] F. Nufairi, N. Pratiwi, and F. Herlando, "Sentiment analysis of Threads application reviews on Google Play Store using support vector machine algorithm," JIPI, vol. 9, no. 1, pp. 339–348, Feb. 2024, doi: 10.29100/jipi.v9i1.4929.

[9] L. Pan, C.-W. Hang, A. Sil, and S. Potdar, "Improved text classification via contrastive adversarial training," AAAI, vol. 36, no. 10, pp. 11130–11138, Jun. 2022, doi: 10.1609/aaai.v36i10.21362.

[10] S. Kierszbaum, T. Klein, and L. Lapasset, "ASRS-CMFS vs. RoBERTa: Comparing two pre-trained language models to predict anomalies in aviation occurrence reports with a low volume of in-domain data available," Aerospace, vol. 9, no. 10, p. 591, Oct. 2022, doi: 10.3390/aerospace9100591.

[11] J. Dai, H. Yan, T. Sun, P. Liu, and X. Qiu, "Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa," arXiv preprint, arXiv:2104.04986, Apr. 2021, doi: 10.48550/arXiv.2104.04986.

[12] C.A. Deagusti, "Sentiment Analysis of the Development Plan of Indonesia's New Capital City (IKN) Based on Twitter (X) Using the Hybrid RoBERTa-GRU Method," Ph.D. dissertation, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia 2024.

[13] N.A.R. Putri and Ardiansyah, "Sentiment analysis of artificial intelligence progress in Indonesia using BERT and RoBERTa," J. Sci. Informatics, vol. 9, no. 2, pp. 136–145, Nov. 2023, doi: 10.34128/jsi.v9i2.649.

[14] T.L. Anggana, "Sentiment analysis using RoBERTa on user reviews of the National Health Insurance (JKN) mobile application," Doctoral dissertation, UIN Sunan Gunung Djati Bandung, 2024. Accessed: Mar. 29, 2025. [Online]. Available: https://digilib.uinsgd.ac.id/99004/

[15] R.M.R.W.P. Kusuma and W. Yustanti, "Customer review sentiment analysis of the Ruang Guru application using the BERT method," J. Emerg. Inf. Syst. Bus. Intel., vol. 2, no. 3, Jul. 2021, doi: 10.26740/jeisbi.v2i3.41567.

[16] G.S. Al-Husna, D. Asmarajati, I.A. Ihsannuddin, and R. Mahmudati, "Comparison of Naive Bayes and support vector machine methods for sentiment analysis on LinkedIn application user reviews," Storage, vol. 3, no. 2, pp. 139–144, 2024, doi: 10.55123/storage.v3i2.3602.

[17] A.K. Dewi, U.S. Semarang, J.L. T. Lomba, J. Mugas, and S. Semarang, "Sentiment analysis of Sicepat expeditions from Google Play reviews using the Naïve Bayes algorithm," J. Inf. Eng. Inf. Syst., vol. 9, no. 2, pp. 796–805, Jun. 2022, doi: 10.35870/jtik.v8i2.1580.

[18] D.F. Sjoraida, B. Wibawa, K. Guna, and D. Yudhakusuma, "Sentiment analysis of the film Dirty Vote using BERT," JTIK, vol. 8, no. 2, pp. 393–404, Mar. 2024, doi: 10.35870/jtik.v8i2.1580.

[19] S.S. Tandiapa and G.C. Rorimpandey, "Sentiment analysis of user reviews on Threads application using lexicon-based method and Naive Bayes classifier," JCM, vol. 3, no. 1, pp. 339–352, Jan. 2024, doi: 10.36312/jcm.v3i1.

[20] R. Ramadhan, "Sentiment analysis on Maxim app reviews on Google Play Store with K-nearest neighbor," JURIKOM, vol. 10, no. 3, pp. 715–724, Jul. 2023. Accessed: Mar. 29, 2025. [Online]. Available: https://repository.uin-suska.ac.id/74467/

[21] J.U.S. Lazuardi and A. Juarna, "Sentiment analysis of Joox application user reviews on Android using the BERT method," Sci. J. Comput. Informatics, vol. 28, no. 3, pp. 251–260, 2023, doi: 10.35760/ik.2023.v28i3.10090.

[22] C.-Z. Liu, Y.-X. Sheng, Z.-Q. Wei, and Y.-Q. Yang, "Research of text classification based on improved TF-IDF algorithm," in Proc. IEEE Int. Conf. Intell. Robot. Control Eng. (IRCE), Aug. 2018, pp. 218–222, doi: 10.1109/IRCE.2018.8492945.

[23] M.A. Java, M. Syafrullah, and F. Teknologi, "Sentiment analysis of user reviews of the Threads application on the Google Play Store using multinomial Naive Bayes and support vector machine," TICOM J. Technol. Inf. Commun., vol. 12, no. 2, 2024, doi: 10.70309/ticom.v12i2.112.

[24] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[25] F.R. Adi Pratama and S.I. Oktora, "Synthetic minority over-sampling technique (SMOTE) for handling imbalanced data in poverty classification," Stat. J. IAOS, vol. 39, no. 1, pp. 233–239, Feb. 2023, doi: 10.3233/SJI-220080.

[26] V. Chandradev, I.M.A.D. Suarjaya, and I.P.A. Bayupati, "Hotel review sentiment analysis using the BERT deep learning method," Buana Inform. J., vol. 14, no. 2, pp. 107–116, Oct. 2023, doi: 10.24002/jbi.v14i02.7244.

[27] F.I. Septian, I.L. Kharisma, H. Hermanto, and K. Kamdan, "Implementation of the Bidirectional Encoder Representations from Transformers (BERT) method for sentiment analysis of Dana application user comments on Instagram," in Proc. TAU SNARS-TEK Nat. Seminar Eng. Technol., vol. 3, no. 1, pp. 201–210, Jan. 2023, doi: 10.47970/snarstek.v2i1.571.

This page is intentionally left blank