ARTICLE

SENTIMENT ANALYSIS OF MYPERTAMINA ON GOOGLE PLAY STORE USING NAÏVE BAYES FOR SUSTAINABLE POLICIES

Engkus[1)*], Eko Prasetyo[2)], Shofa Nurfauziah[3)], Wibi Alya Azahra[4)], Zahra Nanda[5)]
UIN Sunan Gunung Djati Bandung, Indonesia[1)]
Email: engkus@uinsgd.ac.id [1)]*

Article History
Received: 1 Desember 2024 2024
Accepted: 14 Desember 2024
Published: 30 Desember 2024

Abstract

The MyPertamina application is a digital service developed by Pertamina to make it easier for people to carry out vehicle fuel transactions. However, the implementation of this application has generated various responses from users, both in the form of reviews and star ratings on the Google Play Store. User reactions to this application show a variety of opinions, including criticism and appreciation. This research aims to analyze user sentiment towards the MyPertamina application, which is classified into two main categories, namely positive and negative sentiment. The research dataset was obtained through a process of scraping user reviews on the Google Play Store in the period 1 October to 1 December 2024. The data used includes 4812 reviews with label division: ratings 4 and 5 are considered positive sentiment, while ratings 1, 2, and 3 are considered positive sentiment. sentiment.negative Analysis was carried out using the Python programming language via the Google Colab platform. The dataset is divided into 80% training data and 20% test data to build a sentiment classification model. The research results show that the Naive Bayes Classifier algorithm is used to carry out classification with an accuracy level of 78%, precision 75%, recall 99%, and f1-score 86%. This analysis shows that most sentiment towards the app is negative, reflecting user complaints regarding various technical issues with the app. It is hoped that this research can become a basis for improving the MyPertamina application system and provide insight for further research in sentiment classification.

Keywords: Classification Methods, MyPertamina, Naïve Bayes Classifier, Sentiment Analysis, Text Mining

A. INTRODUCTION

The MyPertamina application is a digital service from Pertamina that offers various features, including searching for the nearest Pertamina location, making digital payments while earning loyalty points, as well as a recording system to make monthly petrol purchases easier. This application was created by PT Pertamina as a solution for fuel users to simplify, speed up and secure the transaction process for purchasing vehicle fuel (Wagiswari et al., 2023). MyPertamina is used to sell subsidized fuel oil, namely Pertalite and diesel, which has been implemented since July 1 2022 by PT Pertamina Niaga. President Director of Pertamina Patra Niaga, Alfian Nasution, explained that the distribution of subsidized fuel is one of the mandates given to Pertamina Patra Niaga, as the Commercial & Trading Sub-Holding of PT Pertamina (Persero), to meet the needs of affordable energy for the community, in accordance with Presidential Regulation No. 191/2014 and BPH Migas Decree (SK) No. 4/2020 which regulates the distribution of Solar and Pertalite (Hikmawati, 2022).

Along with the development of user needs, the MyPertamina application is also equipped with additional features designed to improve the user experience. One of them is mapping integration which makes it easier for users to find the nearest gas station with high accuracy, so they no longer need to have difficulty finding a refueling location. In addition, this application allows users to make payments easily and quickly via various methods such as e-wallet, bank transfer and credit card, providing greater flexibility in transactions. Furthermore, the recording system provided in the application allows users to monitor their monthly expenses regarding fuel consumption. The Charging Station feature in this application also functions to find out information on public electric vehicle charging stations (Kharisma & Aesyi, 2023).

With these various features, MyPertamina not only functions as a transaction application, but also as a platform that creates user awareness of the importance of using fuel wisely, as well as providing easy accessibility for people to get the energy they need more practically and efficiently. This application has not been completely well received by the public, with varying responses to government policies implemented through the platform. The policies contained in this application have triggered various reactions, both positive and negative. This can be seen from the many reviews that appear on the Google Play Store, where a number of users provide their comments regarding the experience of using the application and the impact of the policies implemented. To gain a deeper understanding of the public's views and perceptions of the application and related policies, a sentiment analysis of the reviews left by users is needed. By conducting this sentiment analysis, the government can gain clearer insight into how effective the policies that have been implemented are and how they impact user satisfaction. The results of this analysis will provide useful information for policy makers to evaluate and, if necessary, take corrective steps to improve service quality and community satisfaction. In this case, the data to be analyzed is a large number of reviews obtained from the Google Play Store, which will then be processed and analyzed using a predetermined sentiment analysis method.

Naïve Bayes Classifier is a method that relies on probability to classify text. The algorithm is simple but has accurate and decent accuracy. In previous research carried out by (Maria et al., 2023). This research utilized the Naïve Bayes Classifier algorithm and produced an accuracy rate of 78%. With a fairly high level of accuracy, this research chose the Naïve Bayes algorithm to differentiate reviews from various applications. In addition, in research compiled by (Darmawan et al., 2023) They suggested using a larger dataset with a less unequal comparison of positive and negative comments. Therefore, in this research, a wider dataset will be used with a balanced variety of comments between positive and negative. In this research, we decided to use the Naïve Bayes Classifier algorithm as the main method in sentiment analysis, considering its various advantages which have proven to be effective. One of the main reasons for choosing this algorithm is its simplicity; Even though it is relatively easy to understand and apply, this algorithm is still able to produce very effective text classification. Additionally, Naïve Bayes has superior capabilities in handling large text datasets, enabling efficient processing of large amounts of information without sacrificing speed or accuracy.

The main objective of this research is to analyze and understand the sentiment contained in user reviews of the MyPertamina application. These sentiments will be categorized into two main classes, namely positive and negative, which reflect how users respond and rate the application. By mapping user opinions into these two categories, the results obtained from this analysis are expected to provide a clear picture of user satisfaction with the MyPertamina application. The results of this classification can later be used as evaluation material for the government, especially Pertamina, in formulating more appropriate policies and responding well to every input that comes from users. After sentiment data is collected and classification is complete, the next stage is to evaluate the performance of the model used in this research.

This evaluation includes measuring several important parameters, such as accuracy (the level of classification accuracy), precision (how precise the model is in identifying positive or negative classes), recall (the model's ability to capture all existing positive or negative cases), and f1-score (which combines precision and recall into one single value to measure overall performance). Through the results of this research, it is hoped that it can make a significant contribution to the government and PT Pertamina Persero, especially in terms of decision making related to developing or updating the MyPertamina application. With clearer information regarding user sentiment, the government and Pertamina can better understand user needs and expectations, and formulate more appropriate steps and policies to improve application quality and user satisfaction.

B. LITERATURE REVIEW

Application

In the midst of the rapid development of the current digital era, the use of software applications has become very important and fundamental to support various activities, whether in the world of business, education or everyday life. In this increasingly connected world, applications are no longer just tools, but have become the main component that simplifies and improves the quality of various activities carried out by individuals and organizations. Both in the context of professional work, learning processes, and daily social interactions, software applications have a very large role. Along with rapid technological developments, various research and development of software technology continues to be carried out with the aim of creating programs that are not just fulfilling basic user needs, such as completing simple tasks, but also providing more comprehensive and extensive solutions.

One form of significant software technology development is the development of applications that are ready to use and designed to carry out certain functions required by users. These applications are not only designed to perform basic functions, but also allow integration with other applications that have similar or complementary functions, so that they can be used simultaneously to achieve more complex goals. In view (Sutanti et al., 2020), An application can be defined as a program that is ready to be used and designed to carry out certain functions for its users, which are not only limited to one application, but can also involve other applications that support each other to achieve greater goals. Therefore, developing applications that facilitate integration with other applications and enable users to perform multiple functions simultaneously is essential in this ever-evolving digital world.

Teks Mining

In a digital era characterized by an explosion of textual data, where information can be found in various formats and sources, text analysis has become one of the most important approaches to unearth hidden information and provide valuable insights. As the volume and complexity of existing data grows, text processing and analysis becomes a crucial step for understanding patterns, trends and relationships that may not be immediately apparent. One of the methods most often used in textual data analysis is Text Mining. Text Mining, or text mining, can be defined as a process that involves sophisticated techniques to extract knowledge and information from large collections of documents or texts. In this process, users can interact with text data intensively using various available analysis tools and techniques to gain a deeper understanding of the content of the text (Findawati & Rosid, 2020).

Furthermore, the use of Text Mining in the media industry also has a big impact, such as in analyzing news or articles to identify trending topics, measure public sentiment, or evaluate the impact of an event or policy. Therefore, the application of Text Mining really supports smarter, more effective decision making, and is based on structured information from textual data which was previously difficult to analyze manually. With its ability to filter relevant information from large amounts of data, this technology allows various parties to make decisions that are better, faster, and more suited to existing needs and context.

Sentiment Analysis

In a world that is increasingly connected through digital technology, the expression of human opinions, feelings and emotions on various topics has become a very important aspect of social interaction. This is increasingly visible with the existence of social media platforms, online forums and various other online applications that allow people to share their views widely. In this context, to analyze emerging opinion patterns and to understand public perceptions of certain issues, a method known as Sentiment Analysis is used. Sentiment Analysis is an information mining technique (data mining) which aims to explore and extract judgments, opinions, attitudes, feelings and reactions to an entity, whether it is an issue, product, service or a particular problem that is currently trending or of public concern. (Asrumi et al., 2023).

This technique works by processing and analyzing unstructured data, such as text contained in customer reviews, comments on social media, or news articles, to identify whether the sentiment contained therein is positive, negative, or neutral. By leveraging algorithms and natural language processing models, Sentiment Analysis enables the processing of large amounts of information to reveal valuable and in-depth insights. These insights can be used to support strategic decision making in various sectors, such as marketing, where companies can analyze consumer reactions to their products or advertising campaigns; in customer relationship management, to evaluate customer satisfaction and respond to feedback more effectively; in politics, to gauge public opinion toward a particular policy or candidate; to social trend analysis, to understand how society responds to developing social or cultural issues.

Naïve Bayes Algorithm

In the world of machine learning and data analysis, various classification methods are developed to group data based on certain patterns or characteristics. One of the simplest yet very effective methods in this task is the Naive Bayes Classifier. Naive Bayes Classifier is a classification method based on Bayes' Theorem, which relates the probability of an event to the available information. This method relies on probability and statistical principles to classify data into predetermined categories (Mustafa et al., 2018).

The characteristic of the Naive Bayes Classifier is a very strong independence assumption (naive), where each feature in the data is considered independent of each other, regardless of how these features are related to each other in the real world. Although this assumption is often too simplistic, especially in cases where the features are interdependent, the method still performs well in many applications, especially when used for large, complex datasets. The main advantage of Naive Bayes lies in its simplicity in implementation and speed in processing data, which makes it a popular choice in situations where speed and efficiency are essential, such as in text analysis or spam detection. With the ability to handle large datasets quickly, the Naive Bayes Classifier is frequently used in a variety of machine learning applications, from sentiment analysis to pattern recognition and document classification.

C. RESEARCH METHODOLOGY

This research will be carried out through a series of stages, starting with data collection through scraping techniques, followed by a text pre-processing process which includes steps such as case folding, cleansing, tokenizing, stemming, and filtering. Next, word weighting was carried out using the TF-IDF (Term Frequency-Inverse Document Frequency) method, followed by data processing and classification testing using the Naive Bayes Classifier (NBC) algorithm. After that, the classification results will be analyzed further (Alhaqq et al., 2022).

This research will use Google Colab as the main platform and the Python programming language for its implementation. The use of Python in sentiment analysis is supported by various libraries that simplify the calculation and analysis process. Some of the libraries used include pandas, numpy, matplotlib, sklearn, and Sastrawi, which need to be installed first before use. The flow diagram of the research stages can be seen in Figure 1 below.

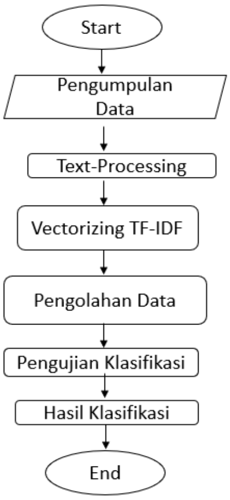


Figure 1. Research Flow
Source: Processed by Researchers, 2024

Data collection

This research collects data through a scraping process from reviews and user ratings of the MyPertamina application on the Google Play Store platform. The data obtained will be used to analyze user sentiment, which is then categorized into two groups, namely positive and negative sentiment. Data collection was carried out in the time period between 1 October to 1 December 2024, using a random sampling method to obtain a representative dataset. Sentiment classification based on ratings is carried out in this way: ratings 4 and 5 are considered positive sentiment, while ratings 1, 2 and 3 are classified as negative sentiment. After the scraping process, the collected data then undergoes a cleansing stage to ensure the quality and cleanliness of the dataset before being analyzed further. The results of the data processing process can be seen in Table 1 below.

Table 1. Example of a Review Dataset

Comment	Value
Most updates...instead of getting easier, they get harder	NEGATIVE
Quite useful	POSITIVE
Registration is very, very difficult	NEGATIVE
Very satisfied with the service of this application	POSITIVE
Something that many state-owned apps usually lack	NEGATIVE

Source: Processed by Researchers, 2024

Text Pre-Processing

Text pre-processing is the first stage in data processing which aims to prepare the data so that it is ready to be used in further analysis, including sentiment analysis. This step is important to ensure that the text data to be analyzed is free from interference or discrepancies that could affect the research results. The pre-processing process includes several stages, such as cleaning (cleaning data), case-folding (uniforming capital letters into lower case letters), tokenizing (breaking text into word units or tokens), stemming (cutting words into their basic form), and filtering (filtering of irrelevant words), all of which were adapted to procedures that have been used in previous studies (Syamsir et al., 2022).

Case-folding

The case-folding stage is carried out by changing all the text in the review to lowercase, so that the writing format becomes consistent and uniform. This step aims to avoid differences between words written in upper and lower case, which actually have the same meaning. In this research, every word in the dataset will be converted to lowercase regardless of whether the beginning of the word is uppercase or lowercase. This ensures that words that have different forms in writing are not treated as different entities in the analysis. An example of applying the case-folding process can be seen in Table 2 below.

Table 2. Casefolding Stages

Before Casefolding	After Casefolding
The government is making it very difficult for people to fill up their vehicles with fuel, they have to use useless applications like this. If they have already registered, they have even been rejected, the reason for the STNK photo is not clear, why is the government implementing something like this.	The government is making it very difficult for people to fill up their vehicles with fuel, they have to use useless applications like this, they have already registered but they have been rejected, the reason for the photo of the vehicle registration is not clear, why is the government implementing something like this.

Source: Processed by Researchers, 2024

Tokenizing

Tokenizing is the stage where each word in a sentence is separated or changed into separate units called tokens. This process aims to break down long texts into smaller chunks that are easier to analyze, so that each word can be treated as a separate entity. By tokenizing, text that was originally a long sentence can be converted into a collection of words or phrases that are easier to manage in subsequent analysis. An example of the tokenizing process can be seen in Table 3 below.

Table 3. Tokenizing Stages

Before Tokenizing	After Tokenizing
The government is making it very difficult for people to fill up their vehicles with fuel, they have to use useless applications like this, they have already registered but they have been rejected, the reason for the photo of the vehicle registration is not clear, why is the government implementing something like this.	'government','very','troublesome','society','to','fill','material' ,'burn','vehicle','just','must','use','' application','which','not','useful','se like','this','already','register','even','rejected','reason','photo',' stnk','no','clear','whatever','why' ,'government','implement','thing','like','this','

Source: Processed by Researchers, 2024

Filtering

Filtering is a stage that aims to remove words that do not have important meaning or relevance in sentiment analysis. This process is carried out to filter out words that do not provide significant additional information, so that the focus of the analysis can be more precisely on relevant words. In this research, the method used is stopwords, namely removing common words that often appear but do not make a significant contribution to understanding sentiment, such as the words "di", "and", "kan", "yang", and so on. By removing these words, the remaining data will be of higher quality and facilitate further analysis. An example of the filtering process can be seen in Table 4 below.

Table 4. Review Filtering Stage

Before Filtering	**After Filtering**
‘ government','very','troublesome', 'society', 'to', 'fill', 'material', 'fuel', 'vehicle', 'just', 'must', 'use', 'app', 'which', 'not', 'useful', 'like', 'this', 'already', 'register', 'even', 'rejected', 'reason', 'photo', 'stnk', 'no', 'clear', 'whatever', 'why', 'government', 'implement', 'thing', 'like this','	‘government’, ‘very’, ‘troublesome’, 'society', 'to', 'fill', 'material', 'fuel', 'vehicle' 'must', 'use', 'app', 'useful','like','already','register' ,'even',rejected','reason','photo','stnk', 'no', 'clear', 'whatever', 'why', 'government', 'implement', 'thing', 'like'

Source: Processed by Researchers, 2024

Stemming

Stemming is a stage that aims to change words that contain affixes (prefixes, suffixes, or inserts) into their basic form, by deleting these affixes. This process is important to unite variations of the same word form, so that words that have similar meanings are considered the same entity in the analysis. In this research, the Sastrawi library is used, which is an Indonesian language processing tool, so that the stemming results obtained will follow the Indonesian grammar rules applied in the library. In this way, each word will be processed according to applicable language rules. An example of the stemming process can be seen in Table 5 below.

Table 5. Stemming Stage

Before Stemming	**After Stemming**
‘government’, ‘very’, ‘troublesome’, 'society', 'to', 'fill', 'material', 'fuel', 'vehicle' 'must', 'use', 'app', 'useful','like','already','register' ,'instead',rejected','reason','photo' ,'stnk' 'no', 'clear', 'whatever', 'why', 'government', 'implement', 'thing', 'like']	‘government’, ‘very’, ‘troublesome’, 'society', 'to', 'fill', 'material', 'fuel', 'vehicle' 'must', 'use', 'app', 'useful','like','already','register' ,'even',rejected','reason','photo','stnk', 'no', 'clear', 'whatever', 'why', 'government', 'implement', 'thing', 'like']

Source: Processed by Researchers, 2024

Vectorizing TF-IDF

The process of giving weight to words in a document is carried out using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizing technique, which aims to measure the importance of a word in the context of the document and the collection of documents as a whole. This technique begins with a pre-processing step to ensure that the text data is ready for further processing. In the TF-IDF method, there are two main components used to determine word weight, namely Term Frequency (TF) and Inverse Document Frequency (IDF).

Term Frequency (TF) measures how often a word appears in a particular document. The more often a word appears, the higher its weight in the document. In this way, words that appear frequently in a document are considered more important or relevant in the context of that document. Meanwhile, Inverse Document Frequency (IDF) is used to measure how unique or specific a word is across a collection of documents. Words that appear very frequently across documents (such as conjunctions or common words) will have low IDF values, because they do not provide much specific information. Conversely, words that rarely appear in many documents will have a higher IDF value, indicating that they are more relevant or specific.

By combining these two components, the TF-IDF method ensures that the most relevant and informative words for analysis are given a higher weight, while words that are too general or do not provide much information are left with a lower weight. This allows the text analysis process to focus more on the words that truly contribute to understanding the context of the document.

Naïve Bayes

This research applies the Naïve Bayes Classifier (NBC) algorithm to process and analyze data. This algorithm is a probability-based classification method, where it uses a statistical model to calculate the possibility of data falling into a certain category or class based on existing attributes or features. Naïve Bayes functions by calculating the probability of each possible class, and selecting the class with the highest probability as the classification result. This algorithm works very efficiently on labeled data, which means data that already has pre-defined categories or classes. Each data is analyzed based on relevant attributes, and based on the combination of these attributes, Naïve Bayes predicts the class of the data. One of the main advantages of Naïve Bayes is its simplicity in calculation, which allows this algorithm to process quickly even on large and complex datasets. Apart from that, Naïve Bayes is also known to have excellent processing speed, making it suitable for use in applications that require real-time analysis.

Classification with Confusion Matrix

Confusion Matrix is an evaluation matrix used to assess the performance of a classification model by comparing the predictions produced by the model with the actual data (correct labels). This matrix presents information about the number of correct and incorrect predictions, divided by category or class in the labeled data. Confusion Matrix is very useful for evaluating how the model works in classifying data into appropriate or incorrect categories.

In classification model evaluation, several evaluation metrics are used to provide a more comprehensive picture of model performance. Some metrics that are often calculated from the results of the Confusion Matrix include:

- Recall: Measures the model's ability to correctly identify positive data, namely the proportion of positive data that is actually detected by the model compared to all existing positive data.

- Precision: Measures how many positive predictions from the model are truly positive, that is, the proportion of positive predictions that are correct compared to all positive predictions made by the model.
- Accuracy: Measures the overall percentage of correct predictions, for both positive and negative categories, compared to the total amount of predicted data.
- F1-Score: This is the harmonic average between precision and recall, which provides an idea of the balance between the two, especially if there is an imbalance between positive and negative classes in the dataset.

By analyzing the Confusion Matrix and these metrics, we can get a clear picture of how well the model performs classification, as well as know the strengths and weaknesses of the model in predicting each category correctly. The results of this evaluation can be found in Table 6 below, which shows the results of calculating evaluation metrics based on this matrix.

Table 6. Confusion Matrix Structure

	Actual Positive (1)	**Actual Negative (0)**
Predicted Positive (1)	TP	FP
Predicted Negative (0)	FN	TN

Source: Processed by Researchers, 2024

Table 6 explains the four main components of the confusion matrix, namely:

- TP (True Positive): The number of correct positive predictions, namely the number of data that is predicted to be positive and is truly positive according to the actual label.
- FP (False Positive): The number of false positive predictions, namely the number of data that are predicted to be positive but are actually negative according to the actual label.
- FN (False Negative): The number of false negative predictions, namely the number of data that are predicted to be negative but are actually positive according to the actual label.
- TN (True Negative): The number of correct negative predictions, namely the number of data that is predicted to be negative and truly negative according to the actual label.

Accuracy is a measure used to determine how well a model can predict the correct value. In simple terms, accuracy describes the comparison between the number of cases predicted correctly (both positive and negative) against all existing data. In other words, accuracy shows how precise the model is in carrying out classification compared to the entire dataset. The formula for calculating accuracy is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision measures the extent to which the model's predictions are accurate in identifying relevant data. It evaluates how many positive predictions actually correspond to positive actual data. The higher the precision value, the fewer false positives made by the model. The formula for calculating precision is:

$$Precision = \frac{TP}{TP\ +FP}$$

Recall measures the extent to which the model can recover relevant information from all the data that should be detected. In this case, recall evaluates the model's ability to identify all truly positive data among all existing positive data. The higher the recall value, the better the model is at capturing all relevant positive data, although it is possible that some negative data will also be detected. The formula for calculating recall is:

$$Recall = \frac{TP}{TP + FN}$$

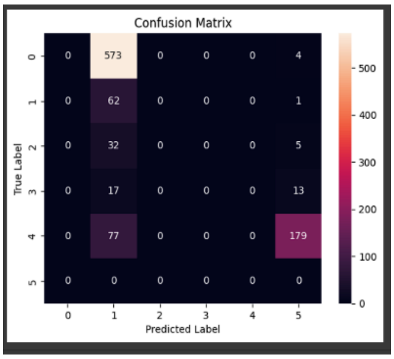
F1-Score is a measure that combines recall and precision by calculating the weighted average of the two metrics. F1-Score provides an overview of the balance between recall and precision, so it is very useful when there is an imbalance between the two, for example in situations where it is important to maintain a balance between capturing as much positive data as possible (recall) and avoiding positive prediction errors (precision).

$$F1\ Score = 2 \text{ x } \frac{Recall \text{ x } Precision}{Recall + Precision}$$

D. RESULT AND DISCUSSIONS

The review data used in this research was obtained through a scraping process from the MyPertamina application on the Google Play Store platform. The scraping process is carried out to collect user review data which can be analyzed to understand the sentiment contained therein. The dataset analyzed comes from a random sample (random sampling) taken in the time period between 1 October 2024 to 1 December 2024. The dataset includes a total of 4812 reviews, which are evenly divided into 2406 negative reviews and 2406 positive reviews, so the distribution of this data is balanced. between the two sentiment categories.

After the data was collected, the dataset was divided into two parts for modeling and evaluation purposes: 80% for training data totaling 3849 reviews, and 20% for testing data totaling 962 reviews. This division aims to train the model using training data, while the test data is used to test the performance of the model that has been trained. This dataset consists of two main columns: the first column contains reviews from users, while the second column contains sentiment labels that categorize each review into two classes: positive or negative, according to the rating or opinion expressed by the user. This dataset sharing and sentiment tagging makes it easier to carry out sentiment analysis and train classification models to predict sentiment in reviews that have not been analyzed previously.



	0	1	2	3	4	5
0	0	573	0	0	0	4
1	0	62	0	0	0	1
2	0	32	0	0	0	5
3	0	17	0	0	0	13
4	0	77	0	0	0	179
5	0	0	0	0	0	0

Figure 2. Confusion Matrix results
Source: Processed by Researchers, 2024

The confusion matrix produced from the classification model provides an overview of the model's performance in predicting data classes. Based on the analysis results, the model was able to classify 573 data from class 0 correctly, but there were 4 class 0 data that were incorrectly classified into other classes. For class 1, the model performance was inadequate because all class 1 data was misclassified, with 62 class 1 data incorrectly identified as class 0. Class 2 also showed similar results, where 32 class 2 data were incorrectly classified as class 0, with no data which was successfully predicted correctly. At class 3, the model showed lower accuracy, with 17 data incorrectly classified as class 0 and another 13 data incorrectly assigned to other classes. Meanwhile, the model showed better performance in class 4, with 179 data successfully classified correctly although there were 77 class 4 data that were incorrectly classified as class 0.

In general, the model performance appears to be biased towards class 0, where most of the data from other classes tends to be classified as class 0. This shows that the model has limitations in recognizing patterns from classes other than class 0, especially in classes 1, 2, and 3 which has a high misclassification rate. Therefore, optimization steps are needed to improve model performance, such as applying dataset balancing techniques to reduce bias, hyperparameter tuning to optimize model parameters, or using a more complex and adaptive classification algorithm. With these improvements, it is hoped that the model can provide more accurate results in classifying data in various classes. This research also evaluates model performance using evaluation metrics obtained from the confusion matrix, such as accuracy, precision, recall, and f1-score. These metrics are very important to measure the accuracy and efficiency of the model in carrying out classification.

- Accuracy describes how correctly the overall model is in classifying data, by calculating the ratio of correct predictions (both positive and negative) to all test data.
- Precision measures the accuracy of a model's positive predictions, namely how many positive predictions are truly positive.
- Recall evaluates the model's ability to capture all relevant positive data among all existing positive data.
- F1-Score combines precision and recall to provide an idea of the balance between the two, especially in cases of class imbalance.

By measuring these metrics, this research provides deeper insight into the model's ability to classify reviews into positive and negative sentiment effectively and efficiently.

Table 7. Confusion Matrix Performance

Performa Confusion Matrix binary class			
Accuracy	**Precision**	**Recall**	**F1-Score**
78 %	75 %	99 %	86 %

Source: Processed by Researchers, 2024

Based on the performance analysis results of the confusion matrix binary class presented in Table 7, the classification model used shows an accuracy of 78%, which means the model is able to classify data correctly at a fairly good level overall. A precision value of 75% indicates that of all the positive predictions made by the model, 75% are correct predictions. This shows the model's ability to avoid misclassification of negative classes as positive classes.

Meanwhile, the very high recall value, namely 99%, indicates that the model is almost able to recognize all positive data correctly. This high recall is an indication that the model has very good sensitivity in detecting positive data, although the potential trade-off with precision still needs to be considered. The F1-score value, which is the harmonic average between precision and recall, was recorded at 86%, which reflects a fairly good balance between prediction accuracy and model sensitivity to positive data.

Overall, the model performance shows satisfactory results for detecting patterns in the data. However, even though the recall value is very high, the lower precision indicates that there is still a number of negative data that are incorrectly classified as positive. To improve model performance, steps such as parameter optimization or adding relevant features can be carried out to improve the balance between precision and recall so that the F1-score value can also be increased. This is important to ensure the model is not only sensitive, but also accurate in its predictions against various classes of data. Overall, these metrics show that the Naïve Bayes Classifier works quite well in classifying user reviews of the MyPertamina application, with quite optimal performance in terms of accuracy, accuracy of positive predictions, and the ability to find relevant positive data.

Based on the confusion matrix analysis, it can be concluded that user reviews of this application show a relatively balanced comparison between positive and negative comments. However, in this study it was found that there were more negative reviews than positive reviews. Several comments indicated that technical problems were one of the main obstacles, such as applications that often crashed, had difficulty logging in, or features that did not function as they should, thus preventing users from achieving their goals. In addition, the user interface (UI) design is considered less user-friendly because it is less intuitive and difficult to navigate, with some features considered too complicated, making it difficult for users to find the functions they need. The registration process is also often complained about, especially because the procedure is considered long and complicated, as well as problems such as failing to receive the OTP or difficulty verifying the account.

Furthermore, the payment features in the application are still considered limited, making it difficult for users to complete transactions. Slow application performance, especially when accessing certain features or making transactions, is also one of the main complaints. On the other hand, many users feel confused due to the lack of clear information regarding how to use the application and applicable policies, which ultimately reduces their satisfaction. In addition, slow response from customer service worsens the user experience, as the problems they encounter are often not addressed quickly, giving the impression that users are not being cared for. Overall these findings suggest that although the app has many positive reviews, significant improvements are needed to improve the user experience, especially in addressing key criticisms expressed through negative reviews.

E. CONCLUSIONS

The use of the Naïve Bayes algorithm in classifying reviews of the MyPertamina application on the Google Play Store shows excellent performance in distinguishing positive and negative reviews. With an accuracy level of up to 78%, this algorithm is able to provide accurate predictions on datasets with a balanced distribution of labels between positive and negative reviews. Additionally, further analysis shows that the Naïve Bayes Classifier (NBC) algorithm has excellent abilities in predicting negative reviews. The dataset used in this research consists of MyPertamina application user review data taken from the Google Play Store. The data distribution is considered quite representative because it includes an almost equal number of positive and negative reviews, thus allowing for more valid analysis results.

The results of this research show that although this application received several positive reviews, the majority of user sentiment was dominated by negative reviews. The main complaints expressed in negative reviews were related to technical problems, such as difficulties in the account registration process and failure to send the OTP code. This problem is a significant obstacle for users, especially when they try to access BBM purchasing services through the application. On the other hand, several users provided positive reviews, most of which highlighted the app's success in limiting access to fuel subsidies to only eligible recipients. This function is considered to help ensure a more targeted distribution of subsidies.

However, to improve the overall user experience, a number of operational improvements to the application system are required. One of the main steps recommended is improving system stability, especially in the registration process and sending OTP codes. This improvement is important to ensure the application can continue to run smoothly, especially when used by many users simultaneously. Apart from that, developers are also advised to carry out regular testing and simulations to identify and overcome potential technical problems before the application is widely used.

For further research development, it is recommended that the performance of the Naïve Bayes algorithm be compared with other algorithms, such as K-Nearest Neighbors (K-NN) or Decision Tree. This aims to evaluate which method is most effective in conducting sentiment analysis on application review data. By comparing different algorithms, research can provide a more comprehensive picture of optimal methods for review classification, thereby supporting better decision making in application management. Future research could also consider using larger and more diverse datasets to increase the generalizability of analysis results.

Recommendation

Practical

Based on the results of sentiment analysis, the main strategic step to increase user satisfaction of the MyPertamina application is to prioritize improvements to the features that are most frequently complained about. Features such as a complicated registration process, OTP code delivery that often fails, and inconsistent application stability should be the main focus in further development. Identification of this problem is done by analyzing negative reviews in depth, so that the root of the problem can be identified and resolved effectively. With these improvements, the user experience can be significantly improved, and user confidence in the application can be restored. In addition, each update must go through rigorous testing to ensure optimal results and avoid the emergence of new problems.

Apart from feature improvements, improving the quality of customer service is also an important priority. The customer service team needs to be strengthened, both in terms of numbers and expertise, to be able to handle complaints quickly and effectively. Implementing a direct feedback system through the application, such as a live chat feature or artificial intelligence-based chatbot, will help speed up problem resolution. This not only makes things easier for users, but also gives the impression that the application is responsive to their needs. Users should feel that their every complaint is heard and taken seriously, so that their satisfaction with the app can increase significantly.

Next, optimizing user experience is another important step that must be taken. In-depth evaluation of the application usage flow, from the registration process to transactions, must be carried out regularly. Processes that are considered too complicated, such as account registration or OTP verification, need to be simplified to make things easier for new users. The application interface design must also be improved to make it more intuitive and easy to understand, so that users do not have difficulty accessing the main features. In addition, developers must ensure that applications can run smoothly on various devices and network conditions, so that technical problems that are often complained about can be minimized.

Finally, transparency in conveying information and establishing good communication with users is also a crucial element in application development. Any policy changes, feature updates, or improvements made must be communicated clearly through effective communication channels, such as social media, email, or discussion forums. This approach not only helps educate users, but also increases their trust in the application. By building open communication relationships, users will feel more appreciated, so that their loyalty to the application can be maintained. Overall, focusing on feature improvements, improving customer service, optimizing user experience, and communication transparency are strategic steps that will have a positive impact on the quality and reputation of the application in the eyes of users.

Further Research

Future research should focus more on sentiment analysis related to service applications, especially applications managed by government agencies. This focus is important because digital transformation in the field of public administration, although much needed, has not yet been explored in depth. Technology-based sentiment analysis, such as those using machine learning methods, including the Naïve Bayes algorithm, can make a significant contribution to understanding people's perceptions of these applications.

This approach not only allows more accurate identification of technical problems and user needs, but also provides useful insights for improving public services. For example, user reviews or complaints can be analyzed to find certain recurring patterns, such as difficulties in accessing services, complaints about the reliability of the application, or criticism of processes that are deemed too complicated. Data obtained from sentiment analysis can be used as a basis for developing policy recommendations that are more responsive to community needs.

REFERENCE

Alhaqq, R. I., Putra, I. M. K., & Ruldeviyani, Y. (2022). Analisis Sentimen Terhadap Penggunaan Aplikasi MySAPK BKN di Google Play Store. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *11*(2), 105–113.

Asrumi, Suharijadi, D., Setiari, A. D., & Wulanda, D. P. (2023). *Analisis Sentimen dan Penggalian Opini* (1st ed.). Eureka Media Aksara.(diunduh tanggal 3 Desember dari Eureka Media Aksara Repository)

Darmawan, G., Alam, S., & Sulistyo, M. I. (2023). Analisis Sentimen Berdasarkan Ulasan Pengguna Aplikasi Mypertamina Pada Google Playstore Menggunakan Metode Naive Bayes. *Jurnal Ilmiah Teknik Dan Ilmu Komputer*, *2*(3), 100–108.

Engkus, E. (2017). Administrasi Publik dalam Perspektif Ekologi. *JISPO Jurnal Ilmu Sosial dan Ilmu Politik*, *7*(1), 91-101.

Engkus, E. (2023). Enhancing Public Services in the West Java Provincial Government: Unraveling Challenges, Defining Essence, and Proposing Solutions. *Journal of Current Social and Political Issues*, *1*(2), 54-61.

Findawati, Y., & Rosid, M. A. (2020). *Buku Ajar Text Mining* (R. Dijaya (ed.); 1st ed.). UMSIDA Press.

Hikmawati, N. K. (2022). Analisis Kualitas Layanan My Pertamina Menggunakan Pendekatan e-GovQual Pada beberapa kota percobaan. *Jurnal Jamika Manajemen Informatika*, *12*(2), 100–111.

Kharisma, K., & Aesyi, U. S. (2023). Analisis Tingkat Kebermanfaatan Mypertamina Menggunakan K-means Clustering. *Jurnal Manajemen Sistem Informasi (JOISM)*, *4*(2), 91–96.

Maria, R., Umayah, R. U., Mahardinny, S., Kalana, D. N., & Saputra, D. D. (2023). Analisis Sentimen Persepsi Masyarakat Terhadap Penggunaan Aplikasi My Pertamina Pada Media Sosial Twitter Menggunakan mEtode Naive Bayes Classifier. *Jurnal Komputer Antartika*, *1*(1), 1–10.

Mustafa, M. S., Ramadhan, M. R., & Thenata, A. P. (2018). Implementasi Data Mining untuk

Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifie. *Creative Information Technology Journal*, *4*(2), 151–162.

Sutanti, A., Komaruddin, M., Mustika, & Damayanti, P. (2020). Rancang Bangun Aplikasi Perpustakaan Keliling Menggunakan Pendekatan Terstruktur. *Jurnal Ilmiah Komputer Dan Informatika*, *9*(1), 1–8.

Syamsir, S., Lutfi, A., Annisa, F. A., Ramadani, I., Putri, N. A., & Nelsi, Y. S. (2022). Efektivitas Penggunaan Aplikasi MyPertamina di Era Kenaikan BBM Bersubsidi. *Prosiding Seminar Nasional Pendidikan, Bahasa, Sastra, Seni, Dan Budaya*, *1*(2), 244–253.

Wagiswari, P. A., Susilawati, I., & Witant, A. (2023). Analisis Sentimen pada komentar aplikasi Mypertamina dengan metode multinomial Naive bayes. *Jurnal ForAI*, *1*(1), 18.