Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.2, December 2024, pp. 531~541
e-ISSN: 2548-6861

Sentiment Analysis on Tabungan Perumahan Rakyat (TAPERA) Program by using Support Vector Machine (SVM)

Rizki Agam Syahputra [1]*, Riski Arifin [2], Suriadi[3]***, Muhammad Iqbal [4]****
* Industrial Engineering Departement, Universitas TeukuUmar
** Industrial Engineering Departement, Universitas Syiah Kuala
** Information Technology Departement , Universitas Teuku Umar
Rizkiagamsyahputra@utu.ac.id [1], riskiarifin@unsyiah.ac.id [2], suriadi@utu.ac.id [3] , iqbal.b20@mhs.usk.ac.id [4]

Article Info

Article history:

Received 2024-10-14
Revised 2024-11-23
Accepted 2024-11-24

Keyword:

Sentiment Analysis,
Support Vector Machine (SVM),
TAPERA.

ABSTRACT

This study aims to analyze public sentiment towards the Housing Savings Program (TAPERA) using the Support Vector Machine (SVM) algorithm. The dataset comprises 16,061 reviews about TAPERA which was gathered from web scrapping and YouTube API. The sentiment analysis results indicate that 99.8% of the reviews are negative, while only 0.2% are positive. The SVM model applied in this study achieved a very high accuracy rate of 99.81%. This indicates that the model is highly effective in classifying sentiments, particularly in identifying negative sentiments. The resulting confusion matrix shows the model's excellent performance in detecting negative sentiments, with no False Positives (FP) and a very high number of True Negatives (TN). However, the model exhibits weaknesses in detecting positive sentiments, as indicated by the presence of several False Negatives (FN) and the absence of True Positives (TP). The findings of this study suggest that the public generally holds a very negative view of the TAPERA program. This insight is crucial for program administrators to consider as they evaluate and improve the program based on negative feedback received from the public. Overall, this research provides important insights into public perceptions of TAPERA and underscores the need for better modeling for more representative sentiment analysis. These findings can serve as a basis for policymakers in designing more effective communication strategies and program improvements to increase public acceptance of TAPERA.

I. PENDAHULUAN

Tabungan Perumahan Rakyat (TAPERA) is a strategic initiative launched by the Indonesian government aimed at providing affordable housing solutions for low- and middle-income citizens in Indonesia. The program's goal is to facilitate home ownership, addressing one of the basic needs of society, and to support social and economic welfare [1]. Given its significance, TAPERA has become a widely discussed topic across various platforms, from social media to discussion forums and online news outlets. Public opinion on the program's effectiveness and implementation reflects collective sentiment toward TAPERA [2].

Understanding public sentiment is crucial in the context of government programs like TAPERA. This sentiment encompasses various views and reactions that can assist the government in evaluating and refining existing policies. Public opinion can be positive, negative, or neutral, depending on individual experiences and perceptions of the program. However, due to the vast volume of data from various sources, manually analyzing public sentiment is impractical and inefficient [3], [4]. Therefore, a systematic and measurable method is needed to process, classify, and evaluate large-scale data with high accuracy. One effective method for this is sentiment analysis using the Support Vector Machine (SVM) algorithm [5], [6][7]. The Support Vector Machine (SVM) algorithm is a well-known machine learning and data mining method for its high accuracy in data classification [8], [9]. Previous research on SVM suggest that SVM works by finding the best hyperplane that can separate

data into two or more classes with maximum margin. In the context of sentiment analysis, SVM can be used to classify public opinion texts into positive, negative, or neutral sentiments. The main advantage of SVM is its ability to handle complex data and find optimal patterns within it. Therefore, SVM is considered an appropriate method for detailed and extensive analysis of public behavior/responses [10], [11].

This research begins with the process of collecting data from various sources containing public opinions or sentiments about the TAPERA program. The data sources for sentiment analysis in this research were obtained by scraping YouTube social media platforms regarding the TAPERA program using Python programming libraries [12]. The collected data consists of text containing opinions and needs further processing for analysis. This data collection process is essential to ensure that the dataset used in this study includes diverse perspectives and is unbiased [3]. The collected data is evaluated using the SVM model to measure negative and positive sentiments found in the social media scraping results. The output is a measurement of positive and negative sentiment regarding the observed issue. To test the model's accuracy, the model is evaluated using a confusion matrix to assess the accuracy of the model's performance [11].

In regard with its purposes, this paper aims to provide significant contributions: for the government and policymakers, the results of sentiment analysis can be used to understand public views on TAPERA and make necessary improvements. This helps in enhancing transparency and effectiveness of government programs. For the public, this research provides a structured channel to express their opinions, and it is hoped that the government can respond more effectively to this feedback. For academics and researchers, this study can serve as a reference for conducting similar studies on other programs or public policies. Additionally, in the field of technology and data science, the development and implementation of the SVM model for sentiment analysis can contribute to the advancement of more sophisticated and efficient text analysis methods. Overall, sentiment analysis of the TAPERA program using the SVM algorithm is an effective and efficient approach to understanding public sentiment on a large scale. This research is expected to provide a clearer picture of public sentiment towards TAPERA and offer valuable insights for the program's improvement in the future. With this sentiment analysis, the government can be more responsive to the needs and aspirations of the public, thereby enhancing overall social and economic welfare.

This research introduces several novel aspects to the field of sentiment analysis and public policy evaluation. Firstly, it combines social media data scraping from a platform as influential and widely used as YouTube, which has not been extensively explored in the context of Indonesian public policy sentiment analysis. Secondly, it implements an SVM model tailored specifically to the linguistic and contextual nuances of the Indonesian language, enhancing the accuracy and relevance of the sentiment classification. Lastly, the integration of sentiment analysis results into actionable insights for policymakers represents a practical application of machine learning that bridges the gap between advanced data science techniques and real-world governmental decision-making. This multi-faceted approach not only contributes to the methodological development in sentiment analysis but also provides a robust framework for future studies aiming to evaluate public opinion on various governmental programs globally. Overall, sentiment analysis of the TAPERA program using the SVM algorithm is an effective and efficient approach to understanding public sentiment on a large scale. This research is expected to provide a clearer picture of public sentiment towards TAPERA and offer valuable insights for the program's improvement in the future. With this sentiment analysis, the government can be more responsive to the needs and aspirations of the public, thereby enhancing overall social and economic welfare.

II. METHODS

This study aims to conduct sentiment analysis on the Public Housing Savings Program (TAPERA) using the Support Vector Machine (SVM) algorithm. The research methodology encompasses several key stages: data collection, data preprocessing, feature extraction, model training and testing, and model performance evaluation. A detailed overview of the analytical system can be seen in the following diagram.

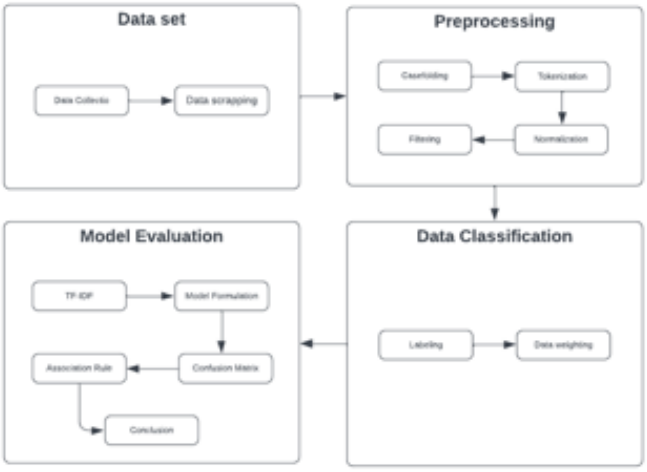

Figure 1. Sentiment analysis model

A. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a robust and effective algorithm for sentiment analysis, a subfield of natural language processing (NLP) that involves determining the sentiment expressed in a piece of text [13]. In sentiment analysis, the goal is to classify text data, such as reviews or social media posts, into sentiment categories like positive, negative, or neutral [13] . Given a training dataset of labeled text documents, SVM aims to find the optimal hyperplane that separates these documents into distinct sentiment classes. The

training data consists of feature vectors xixi derived from text preprocessing and feature extraction techniques, with corresponding sentiment labels yi∈{−1,+1}yi∈{−1,+1}. The SVM optimization problem is formulated as [14]:

$$min_{w,b}\frac{1}{2}||w||^2 \quad subjet\ to\ y_i(\omega.x_i+b)\geq,\forall \qquad (1)$$

This formulation seeks to maximize the margin between the sentiment classes, ensuring better generalization to unseen text data. Text data is often not linearly separable due to the complexity and variability of language. To address this, SVM leverages the kernel trick, which maps the input feature space into a higher-dimensional space where a linear separator can be found. Commonly used kernels in sentiment analysis include the polynomial kernel and the radial basis function (RBF) kernel. The kernel function K(xi,xj)K(xi,xj) allows the SVM to handle non-linear relationships between features. The dual form of the SVM optimization problem, incorporating the kernel function, is [14]:

$$min_{\alpha}\frac{1}{2}\sum_{i,j}a_{i,}a_{j,}y_{i,}y_{j,}k(x_{i,}x_{j})-\sum_{i}a_{i,} \qquad (2)$$

Subject to ∑iαiyi=0 and 0≤αi0≤C, where αi are Lagrange multipliers and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors. In sentiment analysis, feature extraction is crucial for transforming text into numerical vectors suitable for SVM. Common techniques include bag-of-words (BoW), term frequency-inverse document frequency (TF-IDF), and word embeddings like Word2Vec or GloVe. These techniques convert text data into feature vectors xixi_xi that capture the semantic information of the text. By applying SVM with appropriate kernel functions and feature extraction methods, the algorithm can effectively classify text into sentiment categories, even in the presence of complex and non-linear patterns in the data. The use of support vectors, which are critical data points defining the decision boundary, ensures that SVM maintains robustness and high performance in sentiment analysis tasks.

B. Data Set Collection

The dataset for this study was collected through the social media platform YouTube, containing public opinions or sentiments regarding the TAPERA program in 2024. The collected data consists of text (words) obtained using web scraping techniques and the YouTube API (Application Programming Interface) through the Python programming language.

C. Pre-Processing

Following to the data collection process, the next step is data preprocessing sequences as a process to clean and prepare the dataset for analysis. The preprocessing stages include [15].

Casefolding: Converting all letters in the text to lowercase.

Tokenization: Breaking down the text into smaller units called tokens (words or phrases);

Normalization: Standardizing the use of non-standard words.

Filtering: Removing irrelevant components, characters, and punctuation marks from the dataset.

In this study, the data preprocessing was conducted using natural language processing (NLP) libraries with the Python programming language.

D. Data Classification

At this stage, the classification and feature extraction process is conducted on the preprocessed dataset. The classification process involves dividing the dataset into positive, negative, and neutral groups. This classification is performed using the Polarity Score [16]. A Polarity Score is a numerical scale indicating the sentiment or emotional orientation of a text. In sentiment analysis, the Polarity Score can be interpreted as follows:

Negative Score (-1 to 0): shows negative sentiment (not supporting/opposing). The closer the value is to -1, the higher the negative value.

Positive score (0 to 1): shows a positive value (supporting/agreeing). The closer the value is to 1, the stronger the positive sentiment.

In this case, the polarity score and data classification are carried out using the Lexicon library in the Python programming algorithm. The results of the applied algorithm will show the sentiment of each text in the data.

E. Model Evaluation

The model evaluation in this study is performed using TF-IDF weighting and a confusion matrix. TF-IDF (Term Frequency-Inverse Document Frequency) is employed to provide structured weighting to each text in the compiled dataset. In this research, TF-IDF is implemented using the Scikit-learn (SKlearn) library in the Python programming language to convert text data into numerical form. The general formula for TF-IDF is depicted as follows [17]:

$$tf_{t,d} = frequency_{t,d} \qquad (3)$$

$$idf_{t,d} = \ln\left(\frac{1+N}{1+df_t}\right)+1 \qquad (4)$$

$$tf-idf_{td} = idf_{td} * idf_{t,d} \qquad (5)$$

TF-IDF helps in identifying significant words that distinguish between positive and negative sentiments. Additionally, the feature-extracted data is divided into two parts: training data and testing data. The proportion used is 80% for training data and 20% for testing data. The dataset is randomly split to ensure that the training and testing data have a similar distribution. A confusion matrix is an evaluation method for the SVM model to measure the accuracy of the

classification process. In this study, the model used in the confusion matrix is depicted in the following diagram [18-20]:

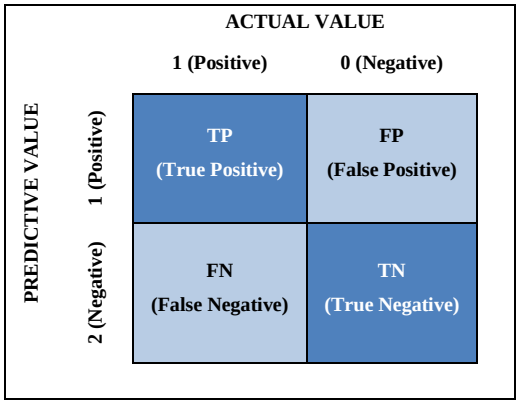

Figure 2. Confusion matrix for SVM model

Where:
True Positive (TP) : The predicted value matches the actual value, or the predicted class matches the actual class.
False Positive (FP) : The predicted value was falsely predicted.
True Negative (TP) : The predicted value matches the actual value, or the predicted class matches the actual class.
False Negative (TP) : The predicted value was falsely predicted.

To calculate the accuracy of the model, the evaluation value of the confusion matrix can be calculated based on the value of Accuracy. The evaluation calculation formula can be seen as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

F. Association Rule

FP-Growth (Frequent Pattern Growth) is one of the algorithms used for discovering association rules in the process of data mining. Association rules are useful for identifying interesting relationships or patterns among items in a large dataset [12]. Simply put, association rules can be interpreted with the function "if ..., then ..." in the context of association, where this method searches for combinations of items in the dataset that meet a minimum support threshold. The support value is determined by the following formula [21].

$$Support\ (A) = \frac{JT(A)}{T}\ x\ 100\% \quad (7)$$

Once frequent patterns with high support have been identified, the association rule value in a dataset that meets association qualifications can be determined using the confidence value (confidence score) for A ⟶ B. The confidence value for A ⟶ B can be calculated using the following formula:

$$Confidence\ P(A/B) = \Sigma\ \frac{Transaction\ A/B}{Transaction\ A}\ x\ 100\% \quad (8)$$

Therefore, equation (8), define P(A | B) as the conditional probability that item A occurs in a transaction given that item B is already present in that transaction. This relationship highlights how the occurrence of items A and B together in transactions is used to calculate the probability of item A appearing in those transactions.

III. Result And Discussion

A. Dataset

The data used in this research consists of a collection of public responses regarding the TAPERA program. The data was gathered using web scraping techniques from the YouTube social media platform. Data collection took place between May 2024 and June 2024. According to the web scraping results, the research successfully collected 16,061 samples, which will be used in the data preprocessing stage. A sample dataset used in this study can be viewed in Table 1.

TABLE 1.
TAPERA DATASET

No	Review
1	Semoga uang tapera ga di korup deh [hoping for no corruption in TAPERA]
2	Potong gaji pejabat !!!!!! [cut goverment official salary]
....	….
10661	UU perampasan Aset = NO PP Tapera = Yes Di awal niat nya baik, gimana cara nya Rakyat punya Rumah. Tapi cara nya yang salah total. Tips rakyat punya Rumah : wujudkan kesejahteraan, sediakan perumahan terjangkau dan bermutu, wajibkan menabung, Larang Pinjol, Judol, dan Paylater.. [Asset Confiscation Law = NO PP Tapera = Yes At the beginning the intentions were good, how do people get a house. But the method is totally wrong. Tips for people to own a house: create prosperity, provide affordable and quality housing, make saving mandatory, prohibit borrowing,online gambling and paylater... }

Based on Table 1, the initial dataset obtained contains various types of data characters formatted astrings (text), including letters, numbers, punctuation marks, emojis, and other specific symbols. Therefore, preprocessing steps are necessary to standardize and remove data/characters that do not contribute meaningful attributes

B. Pre-processing

The process of web scraping resulted in a raw dataset comprising letters, numbers, specific characters, and symbols. To ensure optimal data classification, preprocessing steps are necessary to produce a normalized final dataset that can be accurately interpreted by the system. This involves advanced stages such as case folding, tokenization, normalization, and filtering using the Python programming language. Table 2 presents the preprocessing outcomes for one of the datasets

numbered 16061, serving as a sample in this study. The sample are ranging from unclassified sample, ranging from supportive, neutral and negative sentiment.

Table 2 illustrates a sample of the preprocessing results from the initial dataset, which originally contained punctuation marks and specific characters within the sentences. The preprocessing steps involved removing and altering punctuation marks, converting uppercase letters to lowercase, and transforming words containing affixes into their base forms.

TABLE 2.
PRE-PROCESSED DATA

No	Comment	Case Folding	Tokenizing	Normalization	Filtering	Final Prepocessing
1	Semoga uang tapera ga di korup deh	semoga uang tapera ga di korup deh	[semoga, uang, tapera, ga, di, korup, deh]	[semoga, uang, tapera, enggak, di, korup, deh]	[semoga, uang, tapera, korup, deh]	semoga uang tapera korup
	{Hoping for no corruption in tapera money}	{hoping for no corruption in tapera money}	{hoping, for no corruption in tapera }	{[hoping, money, tapera, no, corrupt, deh]}	{[hoping, money, tapera, no, corrupt, deh]}	{hoping, money, tapera,corrupt}
2	Potong gaji pejabat !!!!!!	potong gaji pejabat	[potong, gaji, pejabat]	[potong, gaji, pejabat]	[potong, gaji, pejabat]	potong gaji pejabat
	{Cut goverment official salary !!!!!!}	{cut goverment official salary}	{[cut, goverment official, salary]}	{[cut, goverment official, salary]}	{[cut, goverment official, salary]}	[cut, goverment official, salary]
...	...	...	...	...	...	...
16060	Pak next request bahas starlink	pak next request bahas starlink	[pak, next, request, bahas, starlink]	[pak, next, request, bahas, starlink]	[next, request, bahas, starlink]	next request bahas starlink
	{Sir request topic on starlink}	{sir request topic on starlink}	{[Sir,request, topic, on starlink]}	{[request, topic, on starlink]}	{[request, topic, on starlink]}	{request, topic, on starlink}
16061	UU perampasan Aset = NO PP Tapera = Yes ...	uu perampasan aset nobrpp tapera yesbrdi awal ...	[uu, perampasan, aset, nobrpp, tapera, yesbrdi...	[uu, perampasan, aset, nobrpp, tapera, yesbrdi...	[uu, perampasan, aset, nobrpp, tapera, yesbrdi...	perampasan aset nobrpp tapera yesbrdi niat rak...
	{Asset Confiscation Law=NO PP Tapera=Yes ...}	{asset confiscation Law=NO PP Tapera=Yes ...}	{[asset confiscation, law, Tapera]}	{[asset confiscation, law, Tapera]}	{[asset confiscation, law, Tapera]}	{[asset confiscation, law, Tapera]}

C. Data Classification And Labelling

At this stage, each normalized dataset will be assigned a label based on the content and meaning of the words within the dataset. Specifically, there are two label groups in the SVM algorithm: positive and negative labels. A positive label indicates data representing positive public feedback, while a negative label indicates negative public feedback on the TAPERA program. To categorize the datasets into sentiment categories, the research first constructs a data dictionary from the dataset used, separating it into a positive data dictionary and a negative data dictionary using the Lexicon library. The results of this dictionary categorization can be seen in the following table 3.

TABLE 3.
SENTIMENT LIBRARY FOR TAPERA ANALYSIS

Sentiment Library	Total
Negative	6610
Positive	3610
Total	10220

Based on the table above, it is identified that a total of 10,220 dictionary entries were formed from the reviews in the research dataset, with 6,610 words having negative connotations and 3,610 words having positive connotations. Using this constructed data dictionary, classification was applied to the entire dataset. The classification results are presented in Table 4.

TABLE 4.
CLASSIFICATION RESULT

No	Polarity Score	Sentiment Categories
1	-5	Negative
2	-5	Negative
3	3	Negative
4	1	Negative
.....	.....	.....
16058	0	Negative
16059	0	Negative
16060	-3	Negative
16061	-9	Negative

Table 4 illustrates a sample of the dataset that has undergone data preprocessing and classification. For instance, data entry number 16061 falls into the negative sentiment category due to having a polarity value of -9. This is because the user's data contains negative words that reject the TAPERA program. Based on this, sentiment categories were summed according to the number of datasets received. The results of this summation and categorization are shown in the following table 5.

TABLE 5.
SENTIMENT CLASSIFICATION SUMMARY

Category	Total
Negative (-)	16036
Positive (+)	25
Total	16061

The aggregate values in Table 5 indicate that out of a total of 16,061 public responses regarding the TAPERA program, there are 16,036 negative sentiments and 25 positive sentiments. To facilitate the analysis process, the percentage of sentiment values for the TAPERA program can be visualized in Figure 3.

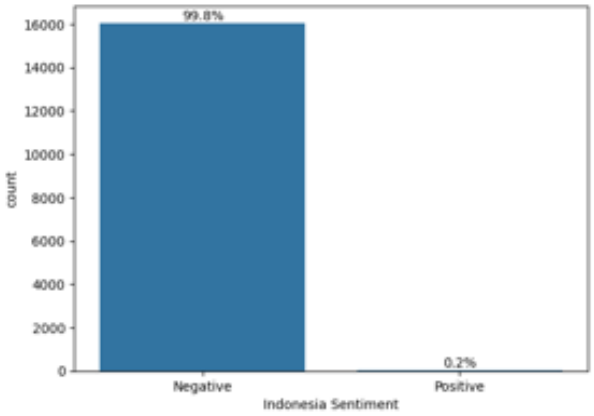

Figure 3. Sentiment value percentage

Based on Figure 3, it is evident that 99.8% of the data falls into the negative sentiment category out of the total dataset (16,036 out of 16,061). Conversely, the analysis of positive sentiment shows that only 0.2% of the responses are positive (25 out of 16,061). In the context of sentiment analysis, this indicates that 99.8% of the public rejects or disagrees with the government's TAPERA program. Conversely, there are very few responses that support the implementation of the TAPERA program.

The sentiment analysis of the TAPERA program shows a highly imbalanced dataset, with 98% of the sentiments classified as negative and only 0.2% as positive. This extreme imbalance raises concerns about potential bias in the model, as it may become overly sensitive to negative sentiments while struggling to accurately classify the few positive ones. Such bias often occurs when a model is exposed to an overwhelming majority class, causing it to prioritize that class at the expense of the minority class. To address this issue, several corrective measures can be implemented. First, data balancing techniques such as oversampling the minority class or undersampling the majority class can help create a more balanced training set. Additionally, evaluation metrics of accuracy evaluation is provided to gain a more nuanced view of model performance, particularly in identifying positive sentiments.

D. Data Modelling

Based on the sentiment classification results, it is found that there is an aggregate total of 98% negative sentiment (rejecting) and 0.2% positive sentiment (agreeing) towards the TAPERA program. To ensure the validity of these results, an evaluation process of the developed model is required. The steps in this stage include data preparation, which involves splitting the dataset into two parts: the testing data and the training data. In this term, the weighting of all data is calculated by using TF-IDF, building the model by selecting the classification algorithm and parameters to be used, and finally evaluating the trained model.

In the data splitting stage, the dataset is divided into two categories: training data and testing data. Training data is the portion of the data required in machine learning classification to learn the characteristics of the data, while testing data is used as input to obtain prediction results. In this study, the dataset is split based on the following ratio analysis [22]:

TABLE 6.
DATA TESTING AND TRAINING RATIO

Ratio	Data separation		Total
	Training	Testing	
60:40	9637	6424	16061
70:30	11245	4818	
80:20	12848	3213	

Therefore, based on Table 6 above, the total number of sentiment data before splitting is 16,06. With a 60:40 data split ratio, there are 9,637 data points for training and 6,424 for testing. At a 70:30 ratio, there are 11,243 data points for training and 4,818 for testing. Lastly, with an 80:20 ratio, there are 12,848 data points for training and 3,213 for testing. Before being used in classification, text data needs to be converted into a numerical form that represents the weight of each word. The results of weighting using TF-IDF are presented in the sample data in Table 7.

TABLE 7.
TF_IDF RESULT

No	Polarity Score	Words	TF Value	IDF Value	TF-IDF Value
1	-5	Semoga [wishing]	0,109	7,892	0,867
2	-5	Deh [only]	0,108	7,892	0,856
3	3	Korup [corrupt]	0,091	7,892	0,718
4	1	Uang [money]	0,068	7,892	0,544
5	-9	tapera [taper]	0,055	7,892	0,439
.....	.....	.....	.....	.....	.....
16057	0	Tolong [help]	0,011	6,976	0,007
16058	0	Sulit [hard]	0,010	6,976	0,007
16059	0	Bang [brother]	0,007	6,976	0,055
16060	-3	Kayak [seems]	0,007	6,976	0,051
16061	-9	Indonesia [Indonesia]	0,007	6,976	0,050

In the classification process, SVM can utilize TF-IDF weighted vector data from training as input to determine the best hyperplane for classifying data into positive and negative classes. Once the model is trained on the training data, TF-IDF weighted vector values from the test data are used as

input to predict the appropriate class for each document. The test data is projected into the same feature space used during training, and then the predicted class is determined based on the position or coordinates of the vector relative to the hyperplane defined during training.

E. SVM Model

To determine the optimal parameters for the classification model, this study utilized the GridSearch method to identify the best combination of kernel and parameter C that enhances the model's performance. The kernel used was the Linear kernel, and parameter C was tested with the values [0.01, 0.1, 1, 100]. Following the optimization process, the optimal combination identified was the Linear kernel with a parameter C value of 0. This parameter set is then used to build the SVM classification model and train it on the training data before evaluating it on the test data. The training and testing of the optimized model are carried out using the following syntax.

```
svm_model = svm.SVC(kernel='linear', C=0)
    svm_model.fit(tf_train, y_train)
 predicted = svm_model.predict(tf_test)
```

The syntax above shows that SVM is employing a linear kernel with a parameter C set to 0. The 'fit' method is then used on the pre-processed and weighted training data. Once the model is trained, it is used to make predictions on new data, specifically the test data, to determine their labels or sentiments. To assess the performance and accuracy of the classification model, an evaluation is carried out in the subsequent step.

F. SVM Model Formulation

Based on the SVM model constructed, the study then measures the accuracy rates for 3 data split ratios as shown in Table 6. The measurement results can be seen in the table 6.

TABLE 8.
SVM ACCURACY

Ratio	Data separation		Accuracy level
	Training	Testing	
60:40	9637	6424	90,51%
70:30	11245	4818	96,27%
80:20	12848	3213	99,81%

Based on model evaluation, the 80:20 ratio is considered the best ratio as it achieved an accuracy of 99.81% compared to other model ratios. An accuracy of 99.81% in an 80:20 split model formulation indicates strong performance of the SVM model in sentiment analysis of TAPERA sentiments. It reflects high predictive accuracy and suggests that the model is effective in classifying sentiments, which is crucial for understanding public perceptions and guiding policy decisions related to the TAPERA program. Therefore, a confusion matrix can be computed to generate classification results for sentiment analysis of the TAPERA program. The results of the confusion matrix can be seen in the following figure 4.

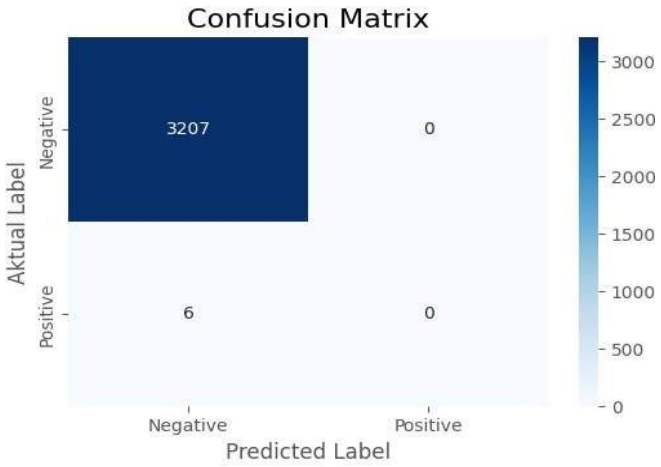

Figure 4. Confusion Matrix

In essence, the detailed explanation of the confusion matrix is as follows:

- True Negative (TN): The number of samples that are actually negative and classified as negative. Here, 3207 negative samples are correctly classified as negative.
- False Positive (FP): The number of samples that are actually negative but classified as positive. Here, there are no negative samples misclassified as positive.
- False Negative (FN): The number of samples that are actually positive but classified as negative. Here, there are 6 positive samples incorrectly classified as negative.
- True Positive (TP): The number of samples that are actually positive and classified as positive. Here, there are no positive samples correctly classified as positive.

Based on this, there is an imbalance between the negative and positive values in the constructed model. This is caused by the multiple and numerous numbers of negative labels in the dataset and the low number of positive labels. The implication that can be built is due to the low proportion of the community rejecting the TAPERA program and the low response of the public supporting the program. To visualize and facilitate the interpretation of the formed data, a Word cloud is used to show the frequency of words that appear most frequently in the analysis, where the larger the word formed, the more often the word is in the dataset. Figure 4 explains the Word cloud formed in the TAPERA sentiment analysis.

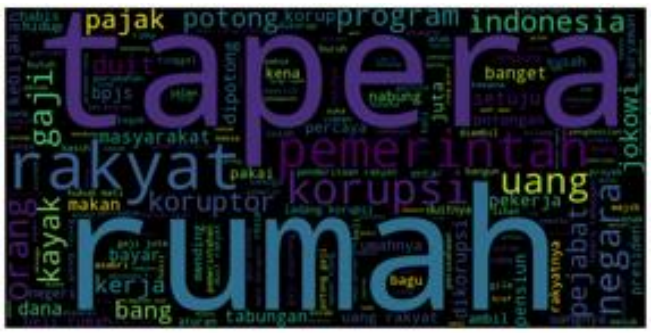

Figure 5. Word cloud for TAPERA

Regarding the most dominant positive words, this study also includes a visualization to display the dominant words based on their frequency of occurrence. Figure 6 illustrates the distribution of positive words in the TAPERA program.

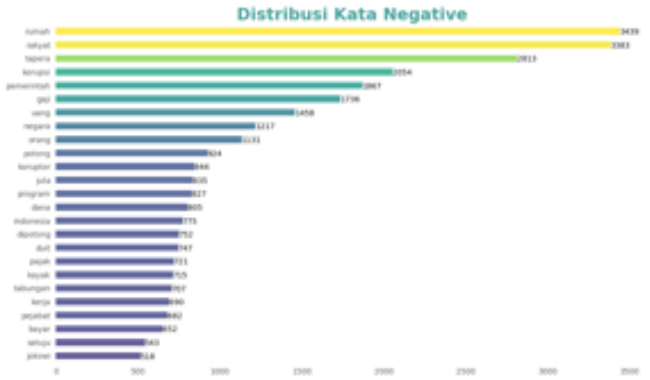

Figure 6. Negative word distribution

Based on the negative word distribution we notice several dominant words that frequently appear in the negative distribution. The most frequently occurring text is 'rumah', which appears 3,439 times. While the next dominants words that also apparent in the negative distribution is include 'Rakyat', 'Tapera', 'Korupsi', 'Pemerintah', and others.

Clarification of the negative distribution words in reviewes sentence include include:

1) Rumahnya aja lo ga jelas di tanah mana, harga dapat berapa, skemanya gaada kenalan gw pns pada ga terima" [“Even the land used for the housing is unclear, the scheme is ambigious, my civil servant friend also never received this program”]: (indicating ambiguity in the TAPERA program)
2) “Kata Lo mulia banget mulia dari Hongkong mulia dari mana rakyat diperas ko dibilang mulia sampah Lo bang” [“ the words that you used is very noble but noble from Hongkong, squeezing the people, calling yourself noble, trash, man”]; (showing differing views and untrust feeling towards the TAPERA program)
3) "Akhirnya selain pajak, Tapera ini target selanjutnya untuk korupsi.” [“finaly apart from tax, now Tapera can be used from curroption source”]; (expressing concerns about corruption in the TAPERA program)

These sentences illustrate users expressing complaints about the TAPERA program. It also highlights shortcomings in TAPERA that need improvement, such as program clarity, indications and opportunities for corruption, and the functionality and success of the TAPERA program when implemented as a comparison, this study also includes a distribution of positive words to gauge public willingness to participate in the TAPERA program. The results of the distribution of positive words in this study can be seen in Figure 7.

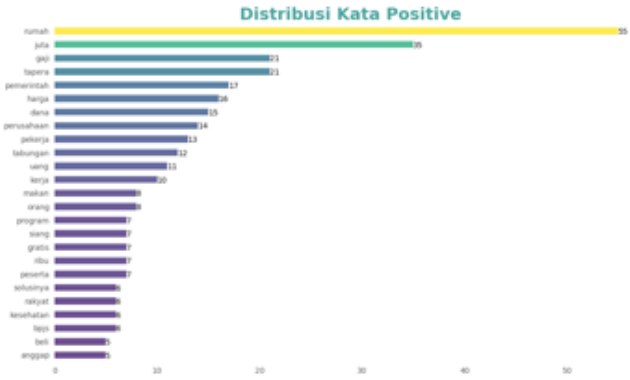

Figure 7. Positive words distribution

The image above displays several dominant words that appear frequently in positive reviews of the TAPERA program. In this context, the most prominent word is 'rumah' (house) appearing 55 times, followed by 'Juta' (million) and 'Gaji' (salary) appearing 21 times. However, it can be observed that the formed positive words have low supportive meaning and cannot represent community responses based on the percentage of positive values.

G. Association Rule

In the context of sentiment analysis and data mining, association rules serve to identify patterns of words or phrases that frequently co-occur in text reviews or comments within a dataset. In this case, association rules are used to observe words that are interrelated when discussing information about the "TAPERA" program. The formulation of association rules concerning the TAPERA program can be found in Table 9.

The table 9 shows the results of association rule analysis using the FP-Growth algorithm to discover relationships between words in public reviews about the TAPERA program. Each row in the table lists combinations of words (antecedents) that frequently appear together with the word "TAPERA" (consequent). From the data, it is evident that words such as "uang" (money), "rakyat" (people), "rumah" (house), and "pemerintah" (government) have a strong correlation with the word "TAPERA". The relatively high support values for these antecedents indicate that these word combinations frequently occur in reviews. Furthermore, the consistent confidence values of 1.0 suggest that whenever an antecedent appears, the consequent "TAPERA" always appears, indicating a close association between these words and the topic of TAPERA in public discourse.

However, despite the high frequency of co-occurrence indicated by confidence and support values, the lift values remaining at 1.0, and leverage and conviction values of 0.0 and inf (infinity), respectively, suggest that this association is no stronger than expected by chance. This implies that while these words frequently appear together with "TAPERA", there is no strong evidence that one word causes the other to appear. This could indicate that these words are often used together in the same context without a strong causal

relationship. Therefore, these findings provide valuable insights for researchers and policymakers to understand patterns in public discourse about TAPERA, but caution is needed in interpreting these results as indications of deeper causal relationships.

TABLE 9.
ASSOCIATION RULES

No	A	B	Sonsequent support	Support	Confidence
0	(uang)	(tapera)	0.11586	1.0	0.116
1	(uang, rakyat)	(tapera)	0.04318	1.0	0.043
2	(uang, rumah, rakyat)	(tapera)	0.01744	1.0	0.017
3	(uang, rumah)	(tapera)	0.04443	1.0	0.044
4	(uang, pemerintah)	(tapera)	0.02782	1.0	0.028
...	...	...	...	...	...
509	(manfaat)	(tapera)	0.01204	1.0	0.012
510	(tolak)	(tapera)	0.01868	1.0	0.019
511	(pengusaha)	(tapera)	0.01162	1.0	0.012
512	(akalan)	(tapera)	0.01121	1.0	0.011
513	(peras)	(tapera)	0.01038	1.0	0.010

H. Sentiment Result

This research examined a dataset containing 16,061 reviews about the TAPERA program. The sentiment analysis revealed a striking imbalance, with 99.8% of the reviews expressing negative sentiments and only 0.2% being positive. This overwhelming negativity suggests significant public dissatisfaction or scepticism towards the program. Such a distribution underscores the critical need for the program's administrators to delve into the underlying causes of these negative perceptions and address them effectively.

The application of the SVM algorithm for sentiment classification in this research demonstrated exceptional performance, achieving an impressive accuracy rate of 99.81%. This high level of accuracy indicates that the SVM model was highly effective in distinguishing between positive and negative sentiments within the dataset. The robustness of the model in handling a heavily skewed dataset further emphasizes the reliability of the findings. However, it also raises questions about the extent of negative sentiment and the specific aspects of the TAPERA program that may be driving such adverse reactions.

The confusion matrix generated from the SVM model provided detailed insights into the model's performance. It showed that the model excelled in identifying negative sentiments, with 3207 True Negatives (TN) and no False Positives (FP). This result highlights the model's precision in correctly classifying reviews that genuinely expressed negative sentiments. However, the matrix also revealed the model's limitations in detecting positive sentiments, as indicated by 6 False Negatives (FN) and no True Positives (TP). This imbalance points to a potential area for improvement, suggesting that the model may need further refinement to better recognize and classify the few positive sentiments present in the dataset. The conclusions drawn from this study are clear: the general public holds predominantly negative views towards the TAPERA program. This widespread negativity has significant implications for the program's future. Program managers and government officials must take these findings seriously, conducting thorough evaluations to understand the specific grievances and concerns of the public. This negative feedback provides a valuable opportunity to implement targeted improvements and policy adjustments that could help mitigate dissatisfaction and enhance the program's acceptance and effectiveness.

Moreover, to make sentiment analysis more representative and comprehensive, it is essential to employ data balancing techniques and further refine the model. By addressing the imbalance in the dataset and enhancing the model's ability to detect positive sentiments, researchers and policymakers can gain a more nuanced understanding of public perceptions. This approach will ensure that both negative and positive feedback is accurately captured and considered, leading to more informed decision-making and better program outcomes for TAPERA. Such improvements in sentiment analysis methodologies will ultimately contribute to a more holistic understanding of community sentiment and foster more effective and responsive program management.

V. CONCLUSION

Based on the study "Sentiment Analysis of the Community Towards the TAPERA Program Using Support Vector Machine (SVM)," it was found that there were 16,061 reviews about TAPERA in the dataset used. The analysis results showed that 99.8% of these reviews had negative sentiments, while only 0.2% were positive. The use of SVM algorithm for sentiment classification in this research demonstrated a very high model accuracy of 99.81%. This high accuracy rate indicates that the SVM model used is capable of classifying sentiments very effectively, despite the majority of reviews tending to be negative. The confusion matrix generated showed that the model performed well in identifying negative sentiments, with 3207 True Negatives (TN) and no False Positives (FP). However, there were limitations in detecting positive sentiments, as evidenced by 6 False Negatives (FN) and no True Positives (TP).

The conclusion of this study indicates that the general public holds negative views towards the TAPERA program. The implications of these findings highlight the need for program managers of TAPERA and the Government to conduct evaluations and improvements based on the negative feedback from the public. Additionally, employing data balancing techniques and further model adjustments could enhance the model's performance in detecting positive sentiments, thereby making sentiment analysis more representative and providing more comprehensive insights into public perceptions of the program.

Future Direction

Sentiment analysis of the TAPERA program using SVM provides a nuanced understanding of public sentiment dynamics, revealing patterns, concerns, and influential factors shaping perceptions. The results and discussions derived from such analysis are instrumental in guiding policymakers, program managers, and stakeholders towards enhancing program acceptance, effectiveness, and alignment with public expectations. By leveraging SVM's analytical power, stakeholders can effectively navigate challenges, capitalize on strengths, and foster positive engagement with the public regarding TAPERA and similar social initiatives. Several insights that can be shared based on the result are:

Insights from sentiment analysis can inform policymakers about public concerns and priorities related to TAPERA. This information can guide policy adjustments, communication strategies, or operational changes to address identified issues and enhance program effectiveness. In this regard, the 99,8% of sentiments are leaning toward to the disagreement to the TAPERA program, which can be informed that the government should reconsider for a comprehensive review and responsive policy adjustments to address public concerns. By leveraging the insights gained from sentiment analysis, policymakers can implement targeted strategies to improve program transparency, engage with the public effectively, address specific grievances, build trust, and ensure the program is accessible and inclusive [23]. These actions can help transform public perception, enhance program effectiveness, and ultimately ensure that TAPERA better serves the needs and expectations of the community.

The sentiment analysis of the TAPERA program reveals several critical areas of concern that require targeted improvements to address the overwhelmingly negative public sentiment. One key issue is the lack of transparency in fund management, with many expressing distrust in how the funds are handled. To address this, the government should enhance transparency by providing detailed, publicly accessible reports on fund usage, ensuring third-party audits, and maintaining strict oversight. This can build public trust and mitigate fears of corruption, a recurring theme in the negative reviews.

Another significant concern is the clarity of program guidelines. Many participants find the eligibility criteria and application processes confusing, which contributes to dissatisfaction. To resolve this, the government should simplify the guidelines, ensure the information is clear and accessible, and provide support via a public outreach campaign, FAQs, and helplines. Additionally, the perception of housing affordability and quality is a major issue, and the program should re-evaluate its offerings to ensure they meet the economic realities of the target audience while maintaining high-quality standards.

Finally, addressing the public's concerns about corruption within TAPERA is essential. Implementing robust anti-corruption measures, such as digitalized processes and independent audits, can ensure accountability. In addition, the government should create a platform where the public can provide ongoing feedback, fostering a transparent, responsive communication channel that reassures participants that their concerns are being heard and acted upon. These targeted actions can improve the program's effectiveness and public perception.

In addition, towards the development of the methodology and result of the research, future studies should explore the use of different sampling strategies and ratios to determine the optimal approach for achieving more accurate and representative sentiment analysis results. Another promising direction is the incorporation of multimodal data sources, such as social media posts, news articles, images, and videos, to enrich sentiment analysis. By combining textual data with visual and auditory content, researchers can gain a more holistic understanding of public sentiment towards TAPERA. Developing and testing multimodal sentiment analysis models will be crucial in capturing nuanced sentiments that may not be evident from text alone. This approach will provide a more comprehensive view of public opinions and concerns. Lastly, advancing natural language processing (NLP) techniques will significantly enhance sentiment analysis capabilities. Utilizing state-of-the-art deep learning models like BERT and GPT can improve the accuracy and depth of sentiment analysis. Future research should compare these advanced models with traditional machine learning approaches to identify the most effective methods for sentiment classification. Additionally, implementing temporal sentiment analysis to track changes in public sentiment over time and analyzing regional and cultural differences will provide valuable insights for policymakers, helping them to tailor the TAPERA program to better meet the needs and expectations of the public.

DAFTAR PUSTAKA

[1] H. G. Putra, E. Fahmi, and K. Taruc, 'Tabungan Perumahan Rakyat (Tapera) dan Penerapannya di DKI Jakarta', Jurnal Muara Sains, Teknologi, Kedokteran dan Ilmu Kesehatan, vol. 3, no. 2, p. 321, Jan. 2020, doi: 10.24912/jmstkik.v3i2.5630.

[2] N. Tania, J. Novienco, and dan Dixon Sanjaya, 'Kajian Hukum Progresif Terhadap Implementasi Produk Tabungan Perumahan Rakyat', Kajian Masalah Hukum dan Pembangunan, 2021, [Online]. Available: www.cnnindonesia.com/nasional/20190903212554-20-427289/

[3] F. Aftab et al., 'A Comprehensive Survey on Sentiment Analysis Techniques', International Journal of Technology, vol. 14, no. 6, pp. 1288–1298, 2023, doi: 10.14716/IJTECH.V14I6.6632.

[4] M. Wankhade, A. C. S. Rao, and C. Kulkarni, 'A survey on sentiment analysis methods, applications, and challenges', Artificial Intelligence Review 2022 55:7, vol. 55, no. 7, pp. 5731–5780, Feb. 2022, doi: 10.1007/S10462-022-10144-1.

[5] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, 'Comparing and combining sentiment analysis methods', COSN 2013 - Proceedings of the 2013 Conference on Online Social Networks, pp. 27–37, 2013, doi: 10.1145/2512938.2512951.

[6] M. F. Fakhrezi, A. F. Rochim, D. Mutiara, and K. Nugraheni, 'Comparison of Sentiment Analysis Methods Based on Accuracy Value Case Study: Twitter Mentions of Academic Article', Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 7, no. 1, pp. 161–167, Feb. 2023, doi: 10.29207/RESTI.V7I1.4767.

[7] R. N. Handayani, 'Optimasi Algoritma Support Vector Machine untuk Analisis Sentimen pada Ulasan Produk Tokopedia Menggunakan PSO', Media Informatika, vol. 20, no. 2, pp. 97–108, Jul. 2021, doi: 10.37595/MEDIAINFO.V20I2.59.

[8] A. Shmilovici, 'Support Vector Machines', Data Mining and Knowledge Discovery Handbook, pp. 257–276, 2005, doi: 10.1007/0-387-25465-X_12.

[9] R. Burbidge and B. Buxton, 'An introduction to support vector machines for data mining', 2001.

[10] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, 'Applications of Support Vector Machine (SVM) Learning in Cancer Genomics', Cancer Genomics Proteomics, vol. 15, no. 1, p. 41, Jan. 2018, doi: 10.21873/CGP.20063.

[11] N. H. Ovirianti, M. Zarlis, and H. Mawengkang, 'Support Vector Machine Using A Classification Algorithm', Sinkron : jurnal dan penelitian teknik informatika, vol. 6, no. 3, pp. 2103–2107, Aug. 2022, doi: 10.33395/SINKRON.V7I3.11597.

[12] Z. Zhang, Z. Liu, and C. Qiao, 'Tendency Mining in Dynamic Association Rules Based on SVM Classifier', 2014.

[13] A. Mat Deris, A. Mohd Zain, and R. Sallehuddin, 'Overview of Support Vector Machine in Modeling Machining Performances', Procedia Eng, vol. 24, pp. 308–312, Jan. 2011, doi: 10.1016/J.PROENG.2011.11.2647.

[14] W. Chu, S. S. Keerthi, and C. J. Ong, 'A general formulation for support vector machines', ICONIP 2002 - Proceedings of the 9th International Conference on Neural Information Processing: Computational Intelligence for the E-Age, vol. 5, pp. 2522–2526, 2002, doi: 10.1109/ICONIP.2002.1201949.

[15] O. Devos, G. Downey, and L. Duponchel, 'Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils', Food Chem, vol. 148, pp. 124–130, 2014, doi: 10.1016/J.FOODCHEM.2013.10.020.

[16] Y. Qi and Z. Shabrina, 'Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach', Soc Netw Anal Min, vol. 13, no. 1, Dec. 2023, doi: 10.1007/S13278-023-01030-X.

[17] P. Kurniawati, R. Y. Fa'rifah, and D. Witarsyah, 'Sentiment Analysis of Maxim Online Transportation App Reviews using Support Vector Machine (SVM) Algorithm', Building of Informatics, Technology and Science (BITS), vol. 5, no. 2, Sep. 2023, doi: 10.47065/bits.v5i2.4265.

[18] M. P. Dwi Cahyo, Widodo, and B. Prasetya Adhi, 'Kinerja Algoritma Support Vector Machine dalam MenentukanKebenaran Informasi Banjir di Twitter', PINTER : Jurnal Pendidikan Teknik Informatika dan Komputer, vol. 3, no. 2, pp. 116–121, Dec. 2019, doi: 10.21009/pinter.3.2.5.

[19] J. Friadi, and D. E. Kurniawan, "Analisis Sentimen Ulasan Wisatawan Terhadap Alun-Alun Kota Batam: Perbandingan Kinerja Metode Naive Bayes dan Support Vector Machine," Jurnal Sistem Informasi Bisnis, vol. 14, no. 4, pp. 403-407, Oct. 2024. https://doi.org/10.21456/vol14iss4pp403-407

[20] D. Vonega, A. Fadila, and D. Kurniawan, "Analisis Sentimen Twitter Terhadap Opini Publik Atas Isu Pencalonan Puan Maharani dalam PILPRES 2024", JAIC, vol. 6, no. 2, pp. 129-135, Nov. 2022.

[21] M. Narvekar and S. F. Syed, 'An optimized algorithm for association rule mining using FP tree', in Procedia Computer Science, Elsevier B.V., 2015, pp. 101–110. doi: 10.1016/j.procs.2015.03.097.

[22] F. O. Widarta and R. A. Syahputra, 'The application of machine learning algorithms for assessing the maturity level of palm fruits as the prominent commodity in the Western-Southern Area of Aceh', Operations Excellence: Journal of Applied Industrial Engineering, vol. 16, no. 1, pp. 56–63, Jun. 2024, doi: 10.22441/OE.2024.V16.I1.102.

[23] D. D. Soeprapto, 'SWOT analysis of BP. Tapera: A public housing savings implementing agency in Indonesia', International Journal of Research in Business and Social Science (2147- 4478), vol. 9, no. 6, pp. 230–243, Oct. 2020, doi: 10.20525/IJRBS.V9I6.900.