SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

Sentiment Analysis of YouTube Comments on the East Java Grant Fund Case Using the Support Vector Machine (SVM) Algorithm

Tirta Arya Bimantoro, Yunianita Rahmawati*, Yulian Findawati, M. Alfan Rosid

Department of Computer Science Program, Universitas Muhammadiyah Sidoarjo, Indonesia

*Corresponding Email: yunianita@umsida.ac.id

Abstract. The phenomenon of corruption in Indonesia, particularly the alleged misuse of grant funds in East Java Province, has triggered various public responses, many of which have been expressed through comments on YouTube. This study aims to categorize and analyze public sentiment on this issue by applying the Support Vector Machine (SVM) algorithm. The data used consists of 1,039 relevant comments collected from several YouTube videos related to the case. The comments were processed through text pre-processing stages, which included case folding, tokenizing, text cleaning, removal of stopwords, stemming, and transformation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. Next, the data was manually labeled into three sentiment categories: positive, negative, and neutral. The data was then divided into training and test data with a ratio of 80:20, and used to train and test the classification model. Evaluation was carried out using accuracy, precision, recall, and f1-score metrics. The results of the study show that the SVM algorithm performs well in classifying public sentiment, so it has the potential to be used as a tool to automatically identify public perceptions of developing social and political issues.

Keywords: Corruption, YouTube, Sentiment Analysis, SVM, TF-IDF

I Introduction

Corruption is a serious problem that is deeply rooted in the Indonesian government system. Although various anti-corruption policies and institutions have been established, corruption practices continue to flourish from the central to regional levels. One case that has come to public attention is the alleged misuse of grant funds by individuals in the East Java Provincial Government. This case has received widespread attention from the public, who have expressed their opinions through various social media platforms, such as YouTube[1].

YouTube is now not only a medium for entertainment, but also a means for the public to openly express their opinions on social and political issues. This phenomenon can be utilized through sentiment analysis to systematically capture public perception. Sentiment analysis is a technique in Natural Language Processing (NLP) that aims to identify the polarity of a text, whether it is positive, negative, or neutral. In the text classification process, the Support Vector Machine (SVM) algorithm is known as one of the most effective and accurate methods[2].

A number of previous studies have shown the success of using SVM in analyzing comments about presidential candidate Prabowo Subianto on YouTube using the SVM method, achieving an accuracy of 85%[1]. Desti Mualfah used SVM to classify comments about the deportation of Ustadz Abdul Somad and demonstrated strong classification performance[3]. Putri and Cahyono combined the use of SVM and Naïve Bayes algorithms in assessing public sentiment towards the quality of public services provided by the DKI Jakarta government and found that SVM was able to capture public perceptions effectively[4].

Meanwhile, analyzing sentiment toward the PT. Timah corruption case using SVM and the SMOTE technique was employed to address the imbalanced data distribution. The results of this study prove that the SVM algorithm continues to show optimal performance even with unbalanced data[5]. Fremmuzar and Baita tested various types of SVM kernel applications to evaluate user sentiment towards Telkomsel services on the Twitter platform and found that kernel selection affects model performance[6].

Sugandi et al. used the Naïve Bayes algorithm to classify comments about BPJS policies. Although the algorithm is quite effective, this study did not discuss analysis of corruption issues[7]. Similarly, Algoritme et al. analyzed comments on beauty vlogger channels using SVM, demonstrating the relevance of this method to various social contexts[2].

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

The research gap lies in the lack of studies focusing on sentiment analysis of public comments on YouTube regarding corruption cases at the provincial level in Indonesia, particularly in East Java. This issue is important because it can reflect the public's real perceptions of corruption practices in the region. Therefore, the research t of this study is to fill this gap by analyzing the sentiment of YouTube comments related to the East Java grant case using the SVM algorithm [8].

2 Method

2.1 Research Design

The study design is the steps that will be taken to determine the research from the initial design to testing. The Support Vector Machine (SVM) algorithm is used to group public sentiment towards alleged grant fund misuse in East Java Province based on comments on YouTube. In this study, the SVM algorithm is used with a Radial Basis Function (RBF) kernel because this kernel is effective in handling high-dimensional text data. In addition, a linear kernel was also tested as a comparison. The stages in this study included a literature review, data collection process, text pre-processing, manual labeling, data balancing using SMOTE, dataset division, model training, and classification performance testing[9].

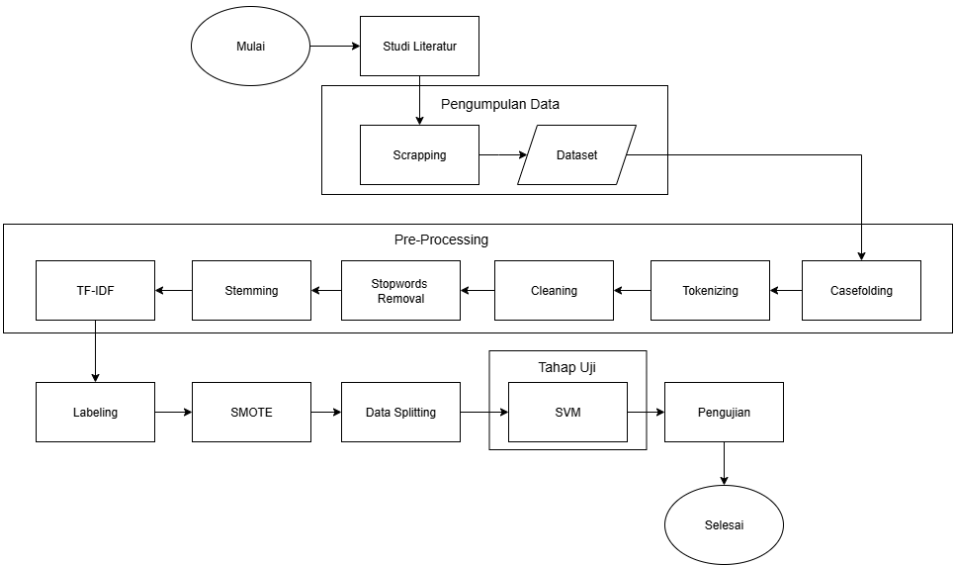


Figure 1. Research Design

2.2 Literature Review

The literature review stage was designed to support scientific references from various sources, including books and scientific articles, as well as video media, to support the completeness and strength of the research foundation. This stage aims to provide a theoretical basis and in-depth understanding of sentiment analysis so that the research has a clear direction and is supported by relevant scientific studies.

2.3 Data Collection

Sources In the data collection process, comments related to the East Java grant fund corruption case were collected from the YouTube platform. The collection was carried out by pulling data using YouTube API version 3 through a Google Console account. Comments were collected from a number of news videos relevant to the research topic, with data taken from 2022 to 2025. After the scraping process was completed, all data was saved in CSV format to facilitate subsequent data processing stages. In total, approximately 1,049 comments were obtained. This dataset was then used in the pre-processing, labeling, and public sentiment classification stages using the SVM method.

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

Table 1. Data Collection

Video URL	Comments
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	East Java only got 9 million, when will it be West Java's turn with 45 million...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	dftr hey Ngawi City and
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	SUBSCRIBE TO CORRUPTION
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Java pest: corruption😂
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	HALAH! What's up with MEGA CORRUPTION AT PERTAMINA? Yeah...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	The KPK is in the wrong place. That stupid governor...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	East Java is full of corruptors😂😂😂 What about the governor?
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Why are there so many problems with grant funds to East Java... ...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Those who once felt safe now can't even sit still...
https://www.youtube.com/watch?v=UqMGrOcC0QM&t=73s	Eradicate all traitors to this country, you monkeys...

2.4 Pre-Processing

Before the data is used for sentiment labeling and classification, a pre-processing stage is carried out to clean and prepare the comment text. This stage aims to convert raw data into a cleaner, more consistent, and structured format so that it can be analyzed optimally by machine learning algorithms such as Support Vector Machine (SVM). This process is very important considering the characteristics of comments on social media, which are often non-standard, full of language variations, symbols, and other non-text elements that can interfere with the accuracy of the analysis. The stages carried out are as follows.

a. Casefolding
   All letters in the comments are converted to lowercase to avoid differences in the spelling of the same word but written with different capitalization.

Table 2. Casefolding

Comments	Casefolding
East Java only 9M, when will it be West Java's turn with 45M...	East Java only 9M, when is it West Java's turn for 45M cum...
dftr hey Ngawi City and	dftr hey Ngawi City and
SUBSCRIBE TO CORRUPTION	Subscription to corruption
Java pest: corruption😂	Java pest: corruption😂
HALAH! What's up with the MEGA CORRUPTION AT PERTAMINA? Yeah...	Oh, come on! What's up with the mega corruption at Pertamina? Yeah...
The KPK is in the wrong place. That stupid governor...	The KPK is in the wrong place. That stupid governor...
East Java is full of corruptors😂😂😂 , what about the governor?	East Java is full of corruptors😂😂😂 What about the governor?
Why are there so many problems with grant funds to East Java?	Why are there so many problems with grant funds to East Java ... ...
Those who once felt safe now cannot even sit still...	Those who once felt safe now cannot even sit still...
Eradicate all traitors to this country, you monkeys...	Eradicate all traitors to this country, you monkeys...

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

b. Tokenizing

This process separates sentences or paragraphs into single word units (*tokens*). Tokenizing is important for breaking down information into basic elements that will be analyzed.

Table 3. Tokenizing

Casefolding	Tokenizing
East Java only 9m when is West Java's turn 45m cum...	[East Java, only, 9m, when, turn, West Java,...
dftr hey Ngawi City and	[dftr, hey, city, Ngawi, and]
subscription to corruption	[*, subscription, corruption, *]
Java pest: corruption😂	[Java, pest, :, corruption😂]
Oh, come on! What's the latest on the mega corruption at Pertamina? Yeah...	[Oh, come on! What's the latest on the mega corruption at Pertamina...
The KPK is barking up the wrong tree. That stupid governor...	[KPK, that, wrong, address, ., governor, that...
East Java is full of corruptors😂😂😂 What about the governor?	[East Java, a hotbed of corruption😂😂😂 , what about the governor...
Why are there so many problems with grant funds to East Java...	[Why, funds, grants, to, East Java, many, problems...
Those who once felt safe now cannot even sit comfortably...	[those, who, used to, feel, safe, now, sitting, even, can't...
Eradicate all traitors to this country, you monkeys...	[eradicate, all, traitors, this country, monkeys...

c. Cleaning

This stage involves removing irrelevant characters or elements in the text, such as punctuation marks, numbers, URLs, mentions (@username), hashtags, emoticons, and other non-alphabetic characters. Cleaning is important so that only core information is processed further.

Table 4. Cleaning

Tokenizing	Cleaning
[East Java, only, 9m, just, when, turn, West Java,...	[East Java, only, m, just, when, turn, West Java, ...
[dftr, hey, city, ngawi, dn]	[dftr, hey, city, ngawi, dn]
[*, subscription, corruption, *]	[subscription, corruption]
[Java, pest, :, corruption😂]	[Java, pest, corruption]
[oh well, !, what, news, mega, corruption, pertamina...	[Oh well, what, news, mega, corruption, Pertamina, ...
[KPK, that, wrong, address, ., governor, sing, ...	[KPK, that, wrong, address, governor, sing, geb..
[East Java, Gae, Bancaan, corruptors😂😂😂 , piye, bu, g...	[East Java, Gae, Bancaan, corruptors, how, ma'am, governor...
[Why, funds, grants, to, East Java, many, problems...	[knp, funds, grants, to, jatim, many, problematic...
[Those who once felt safe now sit uneasily...	[Those who once felt safe now sit, yet...
[eradicate, all, traitors, this country, mon...	[eradicate, all, traitors, this, country, mon...

d. Stopword Removal

Stopword Removal is a text cleaning stage that involves removing words that are general in nature and have no significant influence on the main meaning or context of a text, such as "yang", "di", "dengan", "pada", "untuk". Stopwords are removed so that the model only focuses on words with primary meaning. The Indonesian stopword list can be obtained from NLP libraries such as Sastrawi or NLTK.

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

Table 5. Stopword Removal

Cleaning	Stopword Removal
[East Java, only, m, just, when, turn, West Java, ...	[East Java, m, turn, West Java, m, foundation, achieve...
[dftr, hey, city, ngawi, dn]	[dftr, hey, city, ngawi, dn]
[subscription, corruption]	[subscription, corruption]
[Java, pest, corruption]	[Java, pest, corruption]
[oh well, what, news, mega, corruption, Pertamina, ...	[oh well, news, mega, corruption, Pertamina, dipus...
[KPK, that, wrong, address, governor, sing, geb..	[KPK, wrong, address, governor, sing, geblek, ...
[East Java, Gae, Bancaan, corruptor, how, ma'am, governor...	East Java, Gae, Bancaan, corruptors, how, ma'am, governor...
[Why, funds, grants, to, East Java, many, problematic...	[Why, funds, grants, East Java, problematic, wrong]
[who, used to, feel, safe, now, sit, even, not...	[Those who felt safe, sitting calmly, yes, feel it, especially...
[eradicate, all, traitors, country, this, monkey...	[eradicate, traitors, country, monkeys, dableg,...

e. Stemming

The stemming process converts derivative words into their basic form. For example, the words "menyalahgunakan" (abuse), "penyalahgunaan" (abuse), and "disalahgunakan" (abused) will all be stemmed to "salahguna" (abuse). This technique helps reduce feature dimensions and unify words with the same root. This process is carried out using an Indonesian language stemmer such as Sastrawi.

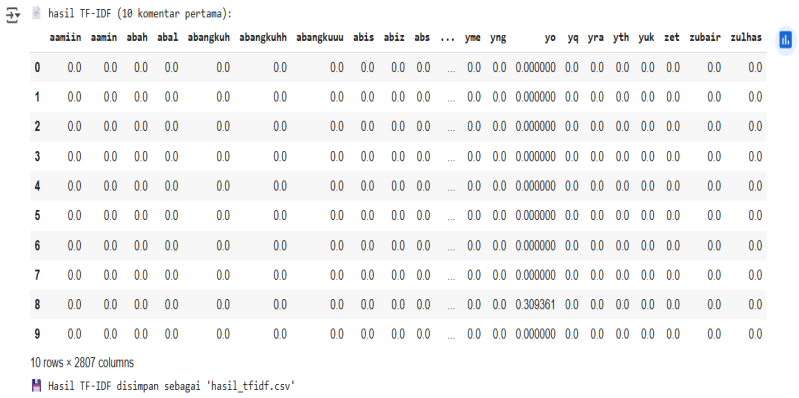
Table 6. Stemming

Stopword Removal	Stemming
[East Java, m, turn, West Java, m, foundation, achieve...	[East Java, m, turn, West Java, m, foundation, achieve, tr...
[dftr, hey, city, ngawi, dn]	[dftr, hey, city, ngawi, dn]
[subscription, corruption]	[subscription, corruption]
[Java, pest, corruption]	[Java, pest, corruption]
[come on, news, mega, corruption, Pertamina, dipus...	[halah, news, mega, corruption, Pertamina, dipus...
[KPK, wrong, address, governor, sing, geblek, ...	[KPK, wrong, address, governor, sing, geblek, ...
[East Java, Gae, Bancaan, corruptor, how, ma'am, governor...	[East Java, Gae, Bancaan, corruptor, how, ma'am, governor...
[KNP, funds, grants, East Java, problematic, wrong]	[KNP, funds, grants, East Java, problems, wrong]
[yg, safe, sit, calm, yo, feel, special, ...	[who, safe, sit, calm, yo, feel, sus, ja...
[eradicate, traitor, country, monkey, dableg,...	[eradicate, traitors, country, monkeys, dableg,...

f. TF-IDF

After the text has been cleaned and stemmed, the next step is to convert the set of words into numerical vectors. The Term Frequency-Inverse Document Frequency (TF-IDF) approach is used to assign high weights to words that appear frequently in a single comment but are rarely found in the entire data set. This strategy aims to identify the words that best represent the content of each document. These TF-IDF values are then used as features in the classification model.

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615



hasil TF-IDF (10 komentar pertama):

	aamiin	aamin	abah	abal	abangkuh	abangkuhh	abangkuuu	abis	abiz	abs	...	yme	yng	yo	yq	yra	yth	yuk	zet	zubair	zulhas
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.309361	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

10 rows × 2807 columns

Hasil TF-IDF disimpan sebagai 'hasil_tfidf.csv'

Figure 2. TF-IDF Results

2.5 Labeling

After pre-processing, each comment is manually grouped into three main sentiment categories, namely positive, negative, and neutral, based on the content and context of the sentence. This process aims to form an accurate and balanced dataset for model training and testing. This labeling technique was also applied by Abdulloh & Pambudi in their study [10].

Table 7. Labeling

Comments	Labeling
East Java only got 9 million, when will it be West Java's turn with 45 million...	-
dftr hey Ngawi City and	-1
CORRUPTION SUBSCRIPTION	-
Java pest: corruption 😂	-1
HALAH! What's up with the MEGA CORRUPTION at PERTAMINA? Yeah...	-
The KPK is in the wrong place. That stupid governor...	-
East Java is full of corruptors 😂 😂 😂 What about the governor?	-1
Why are there so many problems with grant funds to East Java... ...	-1
Those who once felt safe now can't even sit still...	-1
Eradicate all traitors to this country, you monkeys...	-1

2.6 Smote

Initial data shows class distribution imbalance, where negative sentiment is more prevalent than other classes. This can cause the model to lean towards the majority class. To overcome this imbalance, the SMOTE (Synthetic Minority Over-sampling Technique) is used, a method for multiplying synthetic data from minority classes based on the distance between the closest neighbors.
Basic SMOTE formula:

$$x_{\text{baru}} = x_i + \delta \times (x_{zi} - x_i)$$

Explanation:

a. Xi: minority data point
b. Xzi: nearest neighbors of Xi
c. δ: random number between 0 and 1

This technique has been proven effective in improving the performance of classification models on imbalanced data, as described in a study by Asrianto & Herwinanda [11].

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

2.7 Data Splitting

After the SMOTE process, the data is then divided through the data splitting stage:

a. 80% is used for model training (*training data*).
b. 20% is used to test the model's performance (*testing data*).

This division is carried out after the balancing process so that the training data has a proportional sentiment distribution, as also done by Aprilia & Isnain [12].

2.8 Model Evaluation

The classification model is evaluated using four main metrics to measure overall performance, namely: accuracy, precision, recall, and f1-score. This combination of metrics is used to provide a comprehensive assessment of model performance, especially when dealing with unbalanced or noisy data[13]. Based on the test results, the RBF kernel showed more optimal classification performance compared to the linear kernel, so it was used as the main configuration in this study.

Evaluation Formula:

a. Accuracy:

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

b. Precision:

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

c. Recall:

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

d. F1-Score:

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

Description:

a. *TP* = True Positive
b. *TN* = True Negative
c. *FP* = False Positive
d. *FN* = False Negative

3 Results and Discussion

The sentiment classification process for public comments in the case of alleged misuse of grant funds in East Java Province was carried out using the Support Vector Machine (SVM) algorithm. The analysis stage began with an exploration of the initial sentiment distribution, which showed a significant imbalance, with most comments being negative. To overcome this problem, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic data in the minority class so that the distribution between classes would be balanced. Next, after going through the pre-processing stage, the data was divided into two subsets, namely data for training and data for testing. The SVM model was trained using the training data, then its performance was evaluated using a number of metrics, including accuracy, precision, recall, and f1-score, to measure the level of accuracy in classifying comments.

The SVM model evaluation results show excellent classification performance across all sentiment categories, namely positive, neutral, and negative. To illustrate in detail the number of correct and incorrect predictions in each class, a confusion matrix is used. To strengthen the understanding of the comment content, visualization is also carried out with a word cloud that displays the dominant words in each sentiment category. This visualization shows that negative comments are dominated by words such as "corruption" and "money," while neutral comments tend to be informative, and positive comments contain elements of support and hope. Overall, the application of the SVM model combined with data balancing techniques and visual text analysis successfully identified public opinion automatically through digital platforms such as YouTube[14].

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

3.1 Initial Distribution and Data Balancing with SMOTE

The initial distribution of sentiment data shows a very extreme class imbalance, where most comments are classified as negative. Of the total 1,049 comments collected from various relevant YouTube videos, only 3 comments were classified as positive, 30 as neutral, and the remaining 1,016 comments were classified as negative. This kind of imbalance is a common problem in public opinion-based text classification, especially on sensitive or controversial issues such as corruption. This has the potential to cause the classification model to be biased towards the majority class and ignore patterns in the minority class.

To overcome this problem, the SMOTE (Synthetic Minority Over-sampling Technique) was used, which is a popular approach in data balancing. SMOTE works by creating synthetic data examples for minority classes by interpolating from the nearest neighbors in the feature space. This technique is more effective than data duplication because it produces new samples that resemble the distribution of minority classes without repeating the same data[15].

After applying SMOTE, the amount of data in each class became balanced, with 1,016 data points each for positive, neutral, and negative sentiments. This balanced distribution is very important so that the SVM algorithm can learn evenly from each class, improving the model's generalization ability and minimizing classification bias. A visualization comparing the distribution before and after SMOTE is shown in Figure 2.

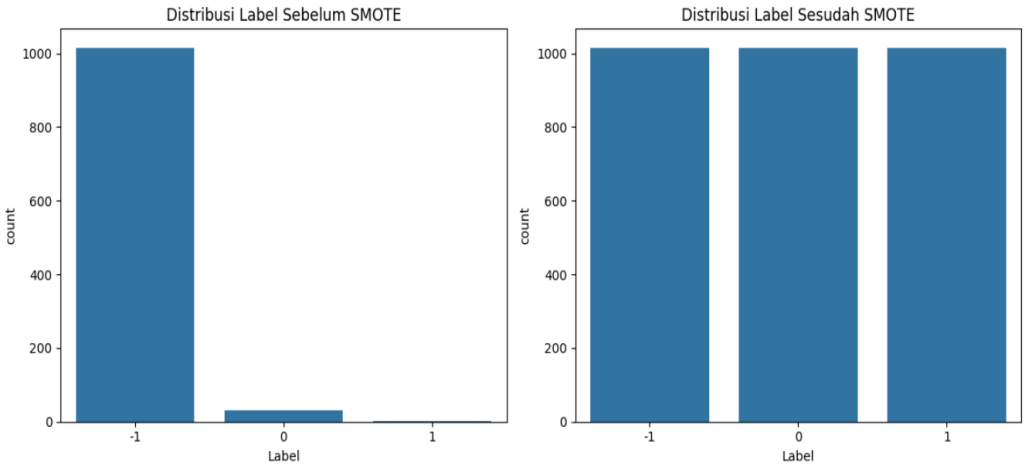


Figure 3. Label Distribution Before and After SMOTE

3.2 SVM Model Performance Evaluation

After the data was balanced, the dataset was divided into two parts: 80% was used for training and 20% for testing. The SVM algorithm is trained using training data with text features that have been processed using the TF-IDF method. Model evaluation is performed using four main metrics: accuracy, precision, recall, and f1-score, as these metrics provide a comprehensive assessment of model performance, especially in text classification with many categories.

Table 8. SVM Model Evaluation Results

Metric	Negative (-1)	Neutral (0)	Positive (1)	Average (Macro avg)
Precision	1.00	0.85	1.00	0.95
Recall	0.83	1.00	1.00	0.94
F1-Score	0.91	0.92	1.00	0.94
Support	203	203	204	610
Accuracy	–	–	–	**0.94 / 94.26**

Based on the table above, it can be concluded that the SVM model has excellent classification performance for all sentiment categories. Although the data was previously imbalanced, the application of SMOTE allowed the model to learn patterns from all three classes proportionally. This is reflected in the high and consistent F1-Score values across

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

all labels. These results show that the model is not only capable of recognizing negative comments, but also sensitive to positive and neutral comments, which are often overlooked by models on imbalanced datasets.

3.3 Confusion Matrix Visualization

To gain a deeper understanding of the classification model's performance, a confusion matrix was used as an evaluation tool. This matrix shows a comparison between the model's predictions and the actual labels, thereby clearly identifying the types of errors that occurred. Based on the visualization shown in Figure 3, the majority of the model's predictions are on the main diagonal, which indicates a high level of accuracy because many predictions match the original labels. Prediction errors such as false positives and false negatives are relatively small and evenly distributed outside the diagonal, indicating that the model is not biased towards a particular class. Thus, the application of the SVM algorithm equipped with the SMOTE data balancing method has proven to be effective in reducing classification errors and producing stable performance across all classes.

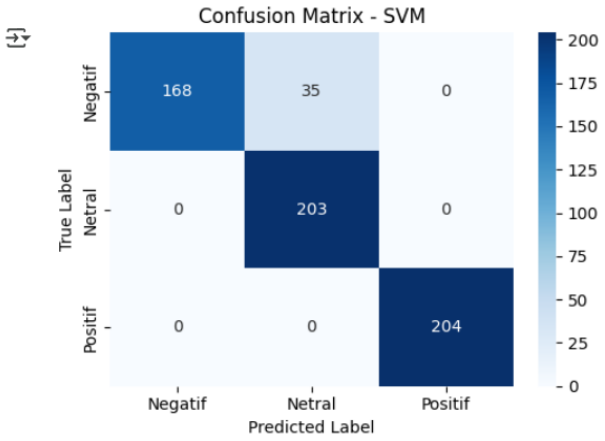


Figure 4. Confusion Matrix for Sentiment Classification Using SVM

3.4 Sentiment Wordcloud Visualization

Wordclouds are used to display the most frequently used words in each sentiment category. This technique is very useful for understanding the content and context of comments without having to read them one by one. Each category is displayed in a separate visualization for easier differentiation, namely Figure 4 for negative comments, Figure 5 for neutral comments, and Figure 6 for positive comments.

Figure 5. Negative Sentiment WordCloud

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

It can be seen that words such as "corruption," "prison," "people," and "money" appear dominantly. The appearance of these words shows that the public has strong and emotional perceptions of the issue of corruption, especially in the context of grant fund misuse. The word "corruption" as the main keyword reflects the public's focus, while the word "prison" indicates the hope that perpetrators will be punished severely. The word "people" indicates that the public feels aggrieved or victimized by these corrupt practices. Meanwhile, the appearance of the word "money" emphasizes that this issue is very sensitive because it concerns public funds that should be used for the public interest.

Figure 6. Neutral Sentiment WordCloud

Displays words such as "case," "report," "grant," and "data." These words are neutral, descriptive, and often appear in comments that only convey information or ask questions without emotional content.

Figure 7. Positive Sentiment WordCloud

Although fewer in number, the word cloud in this category shows words such as "fair," "transparent," "good," and "thank you." This reflects the public's expectations for honest law enforcement and appreciation for those who are considered to be acting correctly.

3.5 Interpretation and Implications

Overall, the findings of this study indicate that the method using Support Vector Machine (SVM) for sentiment classification is effective in analyzing public opinion on sensitive socio-political issues. The application of the SMOTE technique as a data balancing stage is also important in reducing bias towards the majority class. The model is able to accurately recognize patterns of positive and neutral sentiments that were previously in the minority.

Implicitly, this approach can be applied more broadly to monitor public opinion on policy issues, elections, political scandals, and others. Governments or regulatory agencies can use such systems to gather insights from social media automatically and efficiently. In addition, this approach is also relevant in digital social research and the field of political communication.

SaNa: Journal of Blockchain, NFTs and Metaverse Technology
Vol 3, Issue 2, August 2025, Pages 190-201
ISSN: 3030-9832 (Media Online)
DOI: https://doi.org/10.58905/sana.v3i2.615

This research shows that the use of SVM algorithms combined with balancing techniques such as SMOTE and NLP (Natural Language Processing)-based visualization can be an important tool in the era of information openness and digital participation.

4 Conclusion

This study applies the Support Vector Machine (SVM) algorithm in the process of classifying sentiment towards public comments on the YouTube platform related to the alleged corruption of grant funds in East Java Province. The process begins with data collection through YouTube API v3, followed by text pre-processing, manual labeling, data balancing using the SMOTE technique, to the training and performance testing stages of the classification model.

The test results show that the SVM algorithm can identify sentiment with a high accuracy of **94.26**%. The evaluation results indicate that the three main metrics, precision, recall, and f1-score, show consistent and even performance across all sentiment categories, namely positive, neutral, and negative. The SMOTE technique proved to be effective in handling data imbalance, which is a common obstacle in social media-based sentiment analysis. In addition, data visualization through confusion matrices and word clouds provides a deeper understanding of public perception of developing issues.

Overall, this approach proves that a combination of pre-processing techniques, balancing, and the selection of the right algorithm can be used effectively to automatically understand public opinion. It is hoped that this research will assist in the development of sentiment analysis systems in the socio-political domain. In addition, it will serve as a reference for research related to mapping public perception in the digital space.

5 Acknowledgements

With sincerity and respect, the author would like to express his deepest gratitude to all those who have provided encouragement, assistance, and meaningful contributions in every stage of writing this article. All forms of attention and support provided have become an important foundation in completing this article well, namely:

a. Dr. Hidayatulloh, M.Si, as Rector of Muhammadiyah University Sidoarjo, for his support and guidance during the study period.
b. Mr. Irwanto, ST., M.MT., Dean of the Faculty of Science and Technology, for his motivation and valuable academic guidance.
c. Ms. Ade Eviyanti, S.Kom., M.Kom., Head of the Informatics Study Program, for her support and encouragement throughout the course of study.
d. Yunianita Rahmawati, S.Kom., M.Kom., as the academic advisor, for her patience and meticulousness in guiding the author throughout the research process and the preparation of this article.
e. All lecturers in the Computer Science Program who have shared their knowledge, experience, and invaluable inspiration throughout the academic process.
f. The author's mother, for her love, sacrifice, and unceasing prayers. Thank you for being a source of strength, encouragement, and the "main supporter" throughout this challenging process.
g. The author's spouse, who has always been there to provide encouragement, understanding, and emotional support during every difficult phase. Thank you for being both a source of motivation and a thinking partner throughout this lengthy process.
h. All those who cannot be named individually, who have provided assistance, motivation, and support in various forms.

May all forms of support, kindness, and assistance given receive the best reward from Allah SWT. This article would never have been completed without the prayers and encouragement of the extraordinary people surrounding the author.

References

[1] E. Tohidi, R. Perdana Herdiansyah, and E. Wahyudin, "Sentiment Analysis of YouTube Video Comments on the Tvonenews Channel About Presidential Candidate Prabowo Subianto Using Support Vector Machine," 2024. [Online]. Available: Https://Www.Youtube.Com/Watch?V=Wu_Wqhg1ylk

[2] R. T. Adek, Z. Fitri, and S. Chairani Siegar, "Sentiment Analysis of Comments on Indonesian-Language Beauty Vlogger YouTube Channels Using the Support Vector Machine Method," *Jurnal Algoritme*, Vol. 5, No. 2, pp. 164–175, 2025, doi: 10.35957/algoritme.V5i2.9692.

[3] Desti Mualfah, "Sentiment Analysis of YouTube TVOne Comments on Ustadz Abdul Somad's Deportation from Singapore Using the SVM Algorithm," 2023.

[4] R. R. Putri and N. Cahyono, “Sentiment Analysis of Public Comments on DKI Jakarta Government Public Services Using Super Vector Machine and Naive Bayes Algorithms,” 2024.
[5] F. Caroline, R. G. S. Budi, and M. E. Al Rivan, “Sentiment Analysis of the Public Regarding the Corruption Case of PT. Timah Using the Support Vector Machine Method,” *Journal of Computer Science and Informatics*, Vol. 4, No. 1, pp. 43–50, Jul. 2024, Doi: 10.54082/Jiki.141.
[6] P. Fremmuzar and A. Baita, “Kernel SVM Test in Sentiment Analysis of Telkomsel Services on Twitter Social Media,” *Komputika: Journal of Computer Systems*, Vol. 12, No. 2, pp. 57–66, Sep. 2023, Doi: 10.34010/Komputika.V12i2.9460.
[7] Hayati, "Sentiment Analysis of YouTube User Comments on the New Policy of the Social Health Insurance Agency Using Naïve Bayes," 2024.
[8] A. Irma Purnamasari and I. Ali, “Sentiment Analysis of Detik.Com News Comments Using the Support Vector Machine (SVM) Algorithm,” 2024.
[9] N. Kenneth, "The Rise of Corruption Cases in Indonesia Year by Year," *Nathanael Kenneth*, Vol. 2, No. 1, 2024.
[10 F. F. Abdulloh and I. R. Pambudi, "Sentiment Analysis of YouTube Users Towards the Covid-19 Vaccine Program," *Csrid (Computer Science Research And Its Development Journal)*, Vol. 13, No. 3, P. 141, Nov. 2021, Doi: 10.22303/Csrid.13.3.2021.141-148.
[11] R. Asrianto and M. Herwinanda, “Sentiment Analysis of the Increase in Basic Necessities Prices on YouTube Social Media Using the Support Vector Machine Algorithm,” *Coscitech Journal (Computer Science and Information Technology)*, Vol. 3, No. 3, pp. 431–440, Dec. 2022, Doi: 10.37859/Coscitech.V3i3.4368.
[12] Aprilia And A. R. Isnain, “Sentiment Analysis of Twitter Social Media Using the Case of Anti-Corruption Campaigns in Indonesia Using Naive Bayes,” *Jurnal Media Informatika Budidarma*, Vol. 8, No. 2, p. 695, Apr. 2024, Doi: 10.30865/Mib.V8i2.7582.
[13 Muhammad Hilmy Riwanto, “Sentiment Analysis of YouTube Comments Related to the Implementation of Free Nutritious Meals Using the SVM Algorithm Model,” 2025.
[14] L. Rofiqi And M. Akbar, “Sentiment Analysis Related to the Asset Seizure Bill Using Support Vector Machine,” *Jekin - Journal of Informatics Engineering*, Vol. 4, No. 3, pp. 529–538, Aug. 2024, Doi: 10.58794/Jekin.V4i3.824.
[15] H. Hidayat, F. Santoso, and L. F. Lidimillah, “Analysis of YouTube Users’ Sentiment About Rohingya Using the SVM (Support Vector Machine) Algorithm,” *G-Tech: Journal of Applied Technology*, Vol. 8, No. 3, pp. 1729–1738, Jul. 2024, Doi: 10.33379/Gtech.V8i3.4497.