

Hybrid Lexicon and Logistic Regression Sentiment Analysis of YouTube Comments on Suharto's National Hero Designation

Laksmita Dewi Supraba¹, Andi Sunyoto², Setyo³

^{1,3} Department of Management Informatic, Universitas Walisongo Gempol, Pasuruan, Indonesia

² Department of Information Technology, Postgraduate Amikom University, Yogyakarta, Indonesia

Article Info

Article history:

Received April 06, 2026

Revised April 25, 2026

Accepted April 28, 2026

Keywords:

**Sentiment Analysis,
Lexicon-Based Labeling,
Logistic Regression,
YouTube Comments,
Public Opinion,
TF-IDF,
Political Historical.**

ABSTRACT

This study proposes a hybrid approach to sentiment analysis by combining lexicon and Logistic Regression methods based on TF-IDF features, and using Multinomial Naïve Bayes as a benchmark, to classify YouTube comments related to Suharto's appointment as a National Hero. The dataset contains 500 comments TV taken from five YouTube channels: TalkShow tvOne Reload, Kompas Madiun, Tribunnnews, Perspektif, and Tempodotco, 100 comments each. The research stages include text pre-processing, lexicon development, automatic labeling, manual validation of annotations, TF-IDF feature extraction, and sentiment classification. The validation results show a 63% agreement between lexicon labels and manual annotations, with a Cohen's Kappa of 0.15 indicating low inter-annotation agreement. The proposed Logistic Regression framework achieved an overall accuracy of 91%, with a weighted average F1 score of 0.87 and a Macro F1 score of 0.43, slightly better than Naïve Bayes Multinomial, which achieved an accuracy of 90% and a weighted F1 score of 0.85. However, the relatively small performance improvement and high accuracy affected the dominance of the neutral class. The Confusion Matrix for the data shows that the model still has difficulty distinguishing between positive and negative comments the recall for the negative class is 0%, and for the positive class is 20%. Therefore, these results are still preliminary and cannot be generalized.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Laksmita Dewi Supraba,

Departement of Management Informatic, Universitas Walisongo Gempol, Pasuruan, Indonesia

Email: laksmitasupraba@gmail.com

1. INTRODUCTION

Indonesia is a democratic country with the President as head of state and head of government, as stipulated in the 1945 Constitution and Law Number 7 of 2017 [1], [2], [3]. In this democratic system, various policies and political decisions often attract public attention and elicit responses, often expressed through social media, including YouTube.

In political science, the process of selecting and evaluating leaders is explained by the theory of procedural democracy, which emphasizes that elections must be free, fair, and involve public participation [4], [5]. Furthermore, theories of voter behavior and leadership perspectives suggest that public political decisions are influenced by social identity, public perceptions of candidates, and the integrity and capabilities of leaders. One of the most controversial figures in Indonesian political history was President Suharto, who led the country for more than three decades during the New Order era [6].

His leadership was associated with many achievements, but also drew criticism for human rights violations, restrictions on political freedoms, and rampant corruption, collusion, and nepotism. On Heroes' Day on November 10, 2025, President Prabowo Subianto officially awarded the title of National Hero to Suharto based on Presidential Decree of the Republic of Indonesia Number 116/TK/2025 [15]. This decision sparked various public reactions across media outlets, particularly on social media platforms like YouTube, which have become platforms for people to express their

opinions and discuss political and historical issues. Therefore, this study focuses on analyzing public sentiment toward Suharto's designation as a National Hero, as reflected in YouTube comments [7].

Among various social media platforms, YouTube was chosen because it allows users to express more detailed and contextual opinions through the comments section compared to platforms such as X (formerly Twitter), TikTok, Instagram, or online news portals, where comments are often shorter or more limited, as supported by Putri T's (2022) research [8]. Furthermore, YouTube comments remain publicly accessible and directly related to the specific video content, making it a valuable resource for analyzing public opinion on political-historical issues [9].

For example, Mahfudza and Ihksan (2025) analyzed public sentiment towards the 2024 Indonesian Presidential Election using YouTube comments with the Naïve Bayes algorithm and TF-IDF features. The results of this study indicate that YouTube comments are a relevant data source for analyzing public opinion on political issues, especially in Indonesia [10]. Meanwhile, research on Suharto mostly focuses on historical and political aspects, such as inter-state diplomacy during Suharto's reign, as studied by Ahmad Musyalen Firdaus (2023) [11]. In addition, the organizational transformation of the Cabinet Secretariat during the Suharto era, as studied by Riono Widjayanto and Martani (2023), is a study of Suharto's cabinet [12].

Several previous studies have shown that combining lexicon-based methods with machine learning algorithms improves classification performance in various domains [13]. A study conducted by Gifari and Adha (2022) examined Indonesian public sentiment toward film reviews using a combination of TF-IDF and SVM. Their findings confirm that this approach is suitable as a baseline method for sentiment classification tasks in the film review domain. However, the study leaves several gaps: the domain studied is contemporary and does not address historical-political issues; the issue of class inequality has not been adequately addressed; and the unique linguistic characteristics of historically charged commentary, such as ideologically charged diction, have not been examined [14].

Lexicon-based studies show better performance than Naïve Bayes in sentiment analysis of foreign-language texts, with accuracy around 70% higher than Naïve Bayes [15]. The lexicon approach is commonly used for public sentiment analysis research. Lexicon-based sentiment analysis is a more efficient approach than manual methods, thereby potentially supporting research on language behavior analysis.[16]. However, this approach still has limitations in handling complex linguistic forms, such as negation, sarcasm, and speculation. A study by Rahmadhani, Rusanti, and Rosyid (2025) classified 1,493 YouTube comments related to the Indonesian House of Representatives (DPR RI) demonstration using SVM with an Indonesian lexicon and TF-IDF features, achieving 92.4% accuracy [17]. However, this study still has several limitations: the lexicon matches words literally rather than understanding the full meaning of the sentence, the research object is situational and differs from political-historical issues involving collective memory. The use of informal language and slang has not been systematically addressed [18]. These limitations indicate the need to develop sentiment analysis methods for political-historical issues further.

These limitations indicate the need for further improvements in sentiment analysis methods for political and historical issues. Therefore, this study proposes a hybrid approach that combines lexicon-based sentiment labeling with the Logistic Regression algorithm, utilizing TF-IDF features. In addition, this study also compares its performance with Multinomial Naïve Bayes, using the same data and feature representation to assess the performance of sentiment classification on YouTube comments related to the recognition of Suharto as a National Hero.

2. METHOD

2.1 Research Framework

This research framework was implemented using Google Colaboratory with the following workflow:

1. Data Acquisition
2. Data Preparation
3. Lexicon Development
4. Feature Engineering
5. Modeling

6. Evaluation Model

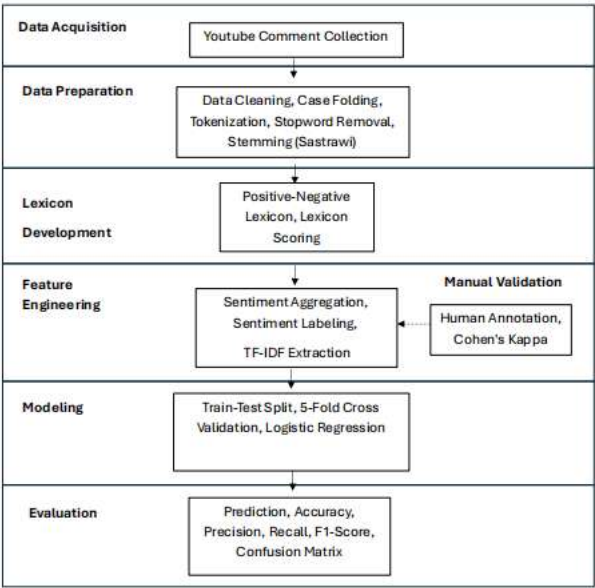


Figure 1. Flow Research Framework

2.2 Data Acquisition

A dataset is a collection of structured data prepared for analysis [19], [20]. The dataset in this study was collected from five YouTube channels: TalkShow tvOne Reload (<https://11nq.com/dzl02ev>), Kompas TV Madiun (<https://11nq.com/g5lh0n2>), Tribunnews (<https://sl1nk.com/556da2d>), Perspektif (<https://11nq.com/o3q9qgt>), and Tempodotco (<https://sl1nk.com/i9hclfr>). Data was automatically retrieved using the Python library youtube-comment-downloader based on the most recent comments, without requiring the YouTube Data API. This library is an open-source, third-party tool available on PyPI/GitHub that extracts YouTube comments without using the official API; details of its internal implementation are beyond the scope of this study's technical contribution, as it was used solely as a data collection instrument.

Comments were collected on November 14, 2025, four days after the government's announcement. The selected videos met four criteria: (1) they specifically discussed Suharto's appointment as a National Hero, (2) they contained views accessible to the public, (3) they provided enough comments for analysis, and (4) they were relevant to the research timeframe. From each channel, only 100 comments that met these criteria were retrieved, yielding a total of 500.

The collected data consists of public comments that can be accessed without special authentication, do not include user personal identities, and are used solely for non-commercial academic purposes. Data collection is limited to 100 comments per channel to minimize server load and mitigate the risk of data misuse. The collected comments are stored in their original form without prior slang normalization. Thus, slang words are not converted to a standard form and are processed according to the characteristics of the original data. The characteristics of the collected videos are presented in Figure 2.

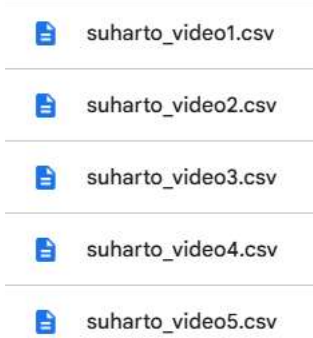


Figure 2. YouTube Comment Extraction Results in Excel Format

The five comment datasets were then combined using the program code in Figure 3, resulting in a combined dataset named data_suharto_gabungan.csv.



Figure 3. Combined YouTube Comments Dataset

The combined dataset consists of 500 comments, all of which were used in the analysis stage without removing any for duplication or empty values.

2.3 Data Preparation

This study uses a hybrid sentiment analysis approach that combines lexicon-based and machine learning methods. The process begins with a data acquisition stage by loading the dataset, followed by a data preparation stage. At this stage, the Comments column containing user comments serves as input to the entire pre-processing process. Next, case folding was applied using `df['case_folding'] = df['Comment'].str.lower()` to convert all characters to lowercase. The resulting text is then cleaned using the `text_preprocessing()` function to remove URLs, web links, mentions, hashtags, numbers, punctuation, and excess spaces, producing a cleaned column [19]. After that, tokenization is performed using `df['tokens'] = df['cleaning'].apply(lambda x: x.split())`, followed by stopwords removal using the built-in stopwords list from the Sastrawi library without any modification.

With this approach, all words in the Sastrawi stopwords list, including negations such as *tidak*, *tak*, and *bukan*, are removed during the stopwords removal stage before the stemming process. Next, stemming is performed using the Sastrawi Stemmer library to obtain the base form of each word as part of the text preprocessing stage [20].

The use of the Sastrawi algorithm is supported by previous research that shows its effectiveness in normalizing affixed words in Indonesian text, thereby reducing morphological variation and improving the quality of text preprocessing [21], [22]. Finally, the stemmed tokens are recombined into a text representation using `df['final_text'] = df['stemming'].apply(lambda word: ' '.join(words))`, which is then used as input for lexicon-based sentiment assessment and TF-IDF feature extraction in the machine learning stage. These preprocessing steps produce clean and structured text data suitable for lexicon construction and subsequent machine learning classification [23].

2.4 Lexicon Development

After data pre-processing, the research proceeded to the lexicon development stage. The lexicon-based sentiment analysis approach is one of the earliest and most widely used methods, assigning each word a polarity value indicating positive, negative, or neutral sentiment. This approach is simple, easy to interpret, and effective in identifying sentiment polarity in text data [15], [16]. In this study, a sentiment dictionary was constructed using two-word sets, `positive_words` and `negative_words`, each containing lists of words with positive or negative polarity, respectively, based on the research context. Next, the `word_score()` function was used to assign a polarity score to each token: +1 for words found in `positive_words`, -1 for words found in `negative_words`, and 0 for words not found in either word set. The scoring process was carried out using `df['token_score'] = df['tokens'].apply(lambda word: [word_score(word) for word in word])` so that each token receives a polarity value based on a predefined lexicon dictionary [26], [27].

Next, the lexicon scoring results are organized into the `lexicon_result` table by separating each token into a single row using the `explode()` function, then renaming the token column to word using `rename()`. The polarity value for each word is calculated using `lexicon_result['word_score'] = lexicon_result['word'].apply(word_score)`. While the polarity table is determined using the `word_label()` function: `lexicon_result['label'] = lexicon_result['word_score'].apply(word_label)`.

Afterward, duplicate data was removed using `drop_duplicates()`, the results were sorted by `word_score` using `sort_values()`, and the index was updated using `reset_index()` to obtain the final lexicon list.

Table 1 displays the Lexicon Score Results generated in this study. The word column contains the processed tokens extracted from YouTube comments; the `word_score` column shows the polarity score assigned by the `word_score()` function; and the label column indicates the classification of each token into the Positive, Negative, or Neutral categories. This information was used to calculate a sentiment score for each comment, followed by TF-IDF feature extraction and classification using the Logistic Regression algorithm.

Table 1. Lexicon Score Results

No	Word	Word_Score	Label
0	Aamin	1	Positif
1	Alhamdulillah	1	Positif
2	Bijaksana	1	Positif

The final sentiment label is determined based on the cumulative score of all tokens in a comment. Comments with a total score greater than zero (>0) are labeled Positive, comments with a total score less than zero (<0) are labeled Negative, and comments with a total score equal to zero ($=0$) are labeled Neutral. This threshold-based labeling mechanism produces simple, easy-to-interpret sentiment labels, which are then used as target classes in the classification process with the Logistic Regression algorithm.

2.5 Feature Engineering

Feature engineering is the process of converting text data into numerical representations that machine learning algorithms can process. In this study, automatically generated sentiment labels from the Lexicon Development stage were used as target classes for classification. The sentiment score for each was calculated using `df['lexicon_score'] = df['token_score']. apply(sum)` and converted into sentiment labels using the `sentiment_label ()` function with `df['label'] = df['lexicon_score']. apply(sentiment_label)` based on predefined threshold rules [24], [25].

To validate the automatic labeling, 100 comments were randomly selected using the `sample ()` function and manually validated by two independent annotators. The level of agreement between annotators was calculated using Cohen's Kappa coefficient via the `cohen_kappa_score ()` function. In contrast, the level of agreement between the automatic labeling results and the first annotator was calculated using Lexicon Agreement [26].

Next, the final preprocessed text (`final_text`) was converted into a numerical representation using the Term Frequency Inverse Document Frequency (TF-IDF) method, which assigns weights to each word based on its frequency of occurrence in the document and its importance to the overall corpus, thus producing a more informative feature representation in the sentiment analysis classification process [27], [28]. This below is formula for TF-IDF:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log \left(\frac{N}{df(t)} \right)$$

Feature extraction was implemented using the default parameters of `TfidfVectorizer ()` via `vectorizer = TfidfVectorizer ()`. The feature matrix was then formed using `X = vectorizer.fit_transform(df['final_text'])`. The sentiment labels generated in the previous stage were stored in the variable `y = df['label']` as the target classes for the classification model's training and evaluation.

2.6 Modeling

The dataset was split into training and test sets at an 80:20 ratio using a stratified train-test split (`stratify=y`) to maintain the proportions of each sentiment class in both sets. The classification process was then performed using the Logistic Regression algorithm as the main model. The model was trained using the training data from the train-test split with L2 regularization configuration, `lbfgs` solver, `class_weight='balanced'`, `C = 1.0`, `max_iter = 1000`, and `random_state = 42`. The `class_weight='balanced'` parameter was applied to reduce the influence of class imbalance on the training process. As a comparison model, this study also implemented the Multinomial Naive Bayes algorithm using the same TF-IDF feature representation. Both models were trained on identical training and test data, allowing direct comparison of their evaluation results.

2.7 Evaluation Model

Evaluation menggunakan metrik accuracy untuk mengukur keseluruhan prediksi yang benar, precision untuk menilai ketepatan prediksi positive, recall untuk melihat kemampuan model menemukan seluruh data positive, serta F1-score sebagai ukuran keseimbangan antara precision dan recall [31],[32].

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Sentiment Distribution

Sentiment analysis was conducted using a dataset of 500 YouTube comments collected from five YouTube videos discussing Suharto's appointment as a National Hero. After data collection, all comments were combined into a single dataset and subsequently used for preprocessing, lexicon-based sentiment labeling, feature extraction, and sentiment classification. The dataset characteristics are presented in Table 2.

Table 2. Dataset Characteristics

Component	Value
Total Comments	500
Data Source	5 YouTube videos

Number of Extracted Files 5 Excel files

This section presents the sentiment distribution of YouTube comments regarding Suharto's appointment as a National Hero using lexicon-based automatic labeling. Each comment was assigned a sentiment score based on the cumulative score of all tokens, then classified into three classes: Positive, Neutral, and Negative, using a threshold-based labeling approach. The sentiment distribution is shown in Table 3

Table 3. Sentiment Distribution

Sentiment	Frequency	Percentage (%)
Positive	24	4.8%
Neutral	451	90.2%
Negative	25	5.0%

3.1.2 Lexicon Validation

To evaluate the quality of the automated labeling, 100 comments were randomly selected for manual validation by two independent annotators. The validation results showed a Cohen's Kappa value of 0.15, indicating a slight agreement between the annotators. Furthermore, the Lexicon Agreement of 63.00% indicates that approximately two-thirds of the labels generated using the lexicon-based approach matched those provided by the first annotator. The relatively low Kappa value indicates that sentiment interpretation in political and historical comments is subjective, potentially leading to differences in assessments among annotators, shown below.

Table 4. Manual Validation Result

Metric	Value
Validation Sample	100 comments
Cohen's Kappa	0.15
Lexicon Agreement	63.00%

Cohen's Kappa value of 0.15 is considered low and, according to Landis and Koch's (1977) interpretation scale, falls into the category of slight agreement (0.00–0.20) [29], [30]. This result is consistent with previous research, which reports that sentiment annotation of political and historical texts is inherently subjective and heavily dependent on contextual interpretation of personal comments [29]. Therefore, differences between annotators are expected when labeling such comments.

3.1.3 Classification Report

In table 5 summarizes the overall performance and the evaluation for each class. Although the model achieved an accuracy of 0.91, this value should be interpreted with caution because the dataset is highly imbalanced and dominated by the Neutral class.

Table 5. Classification Report of Logistic Regression

Class	Precision	Recall	F1-score	Support
Negative	0.00	0.00	0.00	5
Neutral	0.91	1.00	0.95	90
Positive	1.00	0.20	0.33	5
Macro Average	0.64	0.40	0.43	100
Weighted Average	0.87	0.91	0.87	100
Accuracy			0.91	100

The classification report shows that the model performs well on the Neutral class (Recall = 1.00) but fails to detect the Negative class (Recall = 0.00) and shows limited performance on the Positive class (Recall = 0.20). The macro-average F1 score of 0.43 indicates that the model's performance is not evenly distributed across sentiment classes and remains weak for minority classes despite its high overall accuracy.

3.1.4 Cross Validation Results

To test the robustness of the proposed model, five-fold stratified cross-validation was performed during training. The accuracy obtained in each validation fold is presented in Table 6. The results show that the

Logistic Regression model performs relatively consistently across all folds, indicating stability despite the limited dataset size.

Table 6. Five-Fold Cross-Validation Results

Fold	Accuracy
Fold 1	0.96
Fold 2	0.94
Fold 3	0.95
Fold 4	0.93
Fold 5	0.94
Average	0.94
Std Dev	0.01

3.1.5 Confusion Matrix

The classification results on the test data are visualized using a confusion matrix, as shown in the table below.

Table 7. Confusion Matrix

Actual / Predicted	Negative	Neutral	Positive
Negative	0	5	0
Neutral	0	90	0
Positive	0	4	1

Based on the confusion matrix, the model correctly classified 90 Neutral comments and 1 Positive comment, while all Negative comments were misclassified as Neutral.

3.1.6 Majority Class Baseline

As a comparison, this study uses the Majority-Class Baseline, which is a model that always predicts the majority class.

Table 8. Comparison with Majority Class Baseline

Model	Accuracy
Majority-Class Baseline	0.90
Logistic Regression	0.91

3.1.7 Comparison of Logistic Regression with Multinomial Naive Bayes

To evaluate the effectiveness of the proposed model, this study compares the performance of Logistic Regression with Multinomial Naive Bayes using the same test data and evaluation metrics. The comparison results are presented in the Comparison Table of Logistic Regression with Multinomial Naive Bayes. Overall, Logistic Regression performs better than Multinomial Naive Bayes based on accuracy, precision, recall, and F1-score. These results indicate that Logistic Regression is more suitable for classifying sentiment in the dataset used in this study.

Table 9. Comparison of Logistic Regression with Multinomial Naive Bayes

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.91	0.87	0.91	0.87
Multinomial Naive Bayes	0.90	0.81	0.90	0.85

3.1.8 Example Comments

In table 10 presents representative examples of YouTube comments for each sentiment class, including Positive, Neutral, and Negative.

Table 10. Example Comments for Each Sentiment Class

Sentiment	Example Comment

Negative	"Alhamdulillah. Selalu terbayang senyumannya, s..."
Neutral	"Ya otaknya Gunturramli itu otaknya terbalik yg..."
Positive	"Alhamdulillah Pak Soeharto diberi gelar pahlaw..."

3.1.9 Comparison of Lexicon and Human Annotation

The table below presents representative examples comparing sentiment labels generated by lexicon-based approaches with manual annotations. This comparison highlights cases where automated labels differ from human judgments, providing insight into the limitations of lexicon-based methods.

Table 11. Examples of Lexicon and Human Annotation

Comment	Lexicon Label	Human Label	Description
rakyat hrs bersatu besok th 2029 tumbangkan...	Netral	Positif	Political support not captured by the lexicon
D antra pk karno suharto gusdur...	Netral	Positif	Positive opinion expressed without explicit positive lexicon
Megawati makin tua makin pikun...	Netral	Negatif	Implicit criticism toward a political figure
Guntur anda mengatakan PDI-P...	Netral	Negatif	Political criticism not represented in the lexicon

3.2 Discussion

3.2.1 Classification Performance

The Logistic Regression model achieved 91% accuracy with a weighted-average F1-score of 0.87. However, the macro-average F1-score of 0.43 indicates that the model's performance is not uniform across all sentiment classes. The difference between these two metrics indicates that the dominance of the Neutral class influences the high accuracy. Although Logistic Regression performed better than Multinomial Naive Bayes, both models still struggled to classify minority sentiment classes due to the imbalanced data distribution.

3.2.2 Class Imbalance

The data distribution is dominated by the Neutral class, resulting in very limited test data for the Positive and Negative classes and unstable model performance across both. Although Logistic Regression uses `class_weight='balanced'`, the model is still better at recognizing the majority class, as reflected in its high accuracy but low macro-average F1-score. Furthermore, the dominance of the Neutral class is likely influenced by the lexicon-based method's threshold mechanism, which classifies comments that do not meet the threshold for positive or negative sentiment as Neutral.

3.2.3 Confusion Matrix Analysis

The confusion matrix indicates that the model primarily categorizes a significant number of comments as Neutral. All negative comments in the test data were predicted as neutral, while only one positive comment was correctly recognized. This implies that the model's ability to detect patterns is skewed toward the majority class, at the expense of the minority classes. A similar pattern is also seen in the comparison model, Multinomial Naive Bayes, although its overall performance is lower than that of Logistic Regression.

3.2.4 Error Analysis

Several misclassification errors were found in comments containing sarcasm, nostalgia, implicit criticism, or very brief remarks. These characteristics are difficult to identify with lexicon-based approaches because they consider only the polarity (tendency) of words, without understanding their content in the context of the sentence. Furthermore, removing common words (stopwords) and stemming can reduce contextual information, making some implicit or complex comments more difficult to classify.

3.2.5 Lexicon vs Human Labels

Comparison with human annotation shows that lexicon-based labeling still produces discrepancies for comments that are ambiguous, ironic, or mixed in sentiment. The validation results yielded a Lexicon Agreement of 63% and a Cohen's Kappa of 0.15 (slight agreement), indicating that manual validation is still necessary.

3.2.6 Top Terms

The most dominant words in each sentiment class provide an overview of the main issues discussed by YouTube users regarding Suharto's designation as a National Hero. Top terms were obtained based on the highest weighted average Term Frequency–Inverse Document Frequency (TF–IDF) in each sentiment class. The results of the analysis are shown in Table 12.

Table 12. Top Terms of Each Sentiment Class

Rank	Positive	Neutral	Negative
1	alhamdulillah	pahlawan	korupsi
2	hebat	yg	yg
3	pahlawan	pak	negara
4	pak	jadi	jadi
5	nasional	harto	bangsa
6	harto	pki	anak
7	Suharto	suharto	jaman
8	jadi	rakyat	indonesia
9	jasa	gelar	sekarang
10	indonesia	rezi	rezim*

Top terms in the Positive class were dominated by words like "alhamdulillah," "great," "hero," and "service," which conveyed appreciation for Suharto. Conversely, the Negative class was dominated by terms like "korupsi," "negara," "bangsa," "za," and "regime," which reflected criticism of his reign. Meanwhile, the Neutral class was dominated by words such as "hero," "sir," "Suharto," "angkatan," and "rakyat," which were more informative or descriptive.

3.2.7 Political Historical Framing

Thematic analysis shows that positive comments predominantly emphasize narratives about national development, economic stability, and appreciation for Suharto's services. Conversely, negative comments are dominated by discussions about corruption, collusion, nepotism, and human rights violations. Neutral comments generally convey information, discuss, or respond to the process of designating Suharto as a National Hero without any clear evaluative tendencies.

3.2.8 Limitations

This study has several limitations. The data comes from only one YouTube channel, may be influenced by user characteristics and the platform's recommendation algorithm, and cannot be generalized to represent public opinion. The imbalanced class distribution makes it easier for the model to recognize the majority class, leading to lower overall accuracy. Furthermore, some neutral comments may still contain implicit sentiment, sarcasm, or ambiguity that the lexicon-based approach fails to recognize. Because the model was trained using lexicon-based labels, the classification results are more representative of lexicon labeling patterns than independent validation of public sentiment. Therefore, the findings of this study are preliminary and require validation using a larger, more diverse dataset.

4. CONCLUSION

This study proposes a hybrid approach that combines a lexicon-based method with a logistic regression model using TF–IDF features to classify the sentiment of YouTube comments on Suharto's designation as a National Hero. The model achieved 91% accuracy with a weighted-average F1-score of 0.87. However, the macro-average F1-score of 0.43 and the confusion matrix results indicate that the model's performance is not uniform across all sentiment classes. The model demonstrated reliable performance in identifying the Neutral class; however, it struggled to classify most Positive comments correctly and failed to identify any Negative comments in the test dataset. Therefore, overall, these findings position the proposed approach as an initial step toward sentiment classification, one that requires further refinement rather than a finalized solution and is ineffective in detecting minority sentiment classes.

This study used approximately 500 comments from a single YouTube channel, so the results cannot be generalized to represent public opinion. Furthermore, the model was trained using labels generated by the lexicon-based method, so its classification results remain influenced by that method. Dataset size limitations, data source bias, and the presence of sarcasm, irony, and political language can also influence classification results. Therefore, the dominance of the Neutral class cannot be directly interpreted as evidence that most of the public holds a neutral stance on the issue under study.

Further research is recommended using larger datasets from various social media platforms, along with manual annotation as reference labels, the application of class-balanced modeling, transformer-based models such as IndoBERT, and topic modeling and discourse/framing analysis approaches to improve classification quality and provide a more comprehensive understanding of public opinion. This study focused on sentiment analysis of YouTube comments and did not specifically operationalize or analyze the Government Narrative Strategy; therefore, this aspect could be a direction for future research.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided by the Higher Education Service Institution (LLDIKTI) Region VII through the International Reputable Journal Publication Assistance Program.

REFERENCES

- [1] R. Pradhan, "Extracting Sentiments from YouTube Comments," in *Proceedings of the IEEE International Conference on Image Information Processing*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 142–145. doi: 10.1109/ICIIP53038.2021.9702561.
- [2] B. Myagmar, J. Li, and S. Kimura, "Cross-Domain Sentiment Classification with Bidirectional Contextualized Transformer Language Models," *IEEE Access*, vol. 7, pp. 163219–163230, 2019, doi: 10.1109/ACCESS.2019.2952360.
- [3] A. Bondielli and F. Marcelloni, "A survey on fake news and rumor detection techniques," *Inf. Sci. (N. Y.)*, vol. 497, pp. 38–55, Sep. 2019, doi: 10.1016/j.ins.2019.05.035.
- [4] S. K. Lora, I. Jahan, R. Hussain, R. Shahriyar, and A. B. M. A. Al Islam, "A transformer-based generative adversarial learning to detect sarcasm from Bengali text with correct classification of confusing text," *Heliyon*, vol. 9, no. 12, Dec. 2023, doi: 10.1016/j.heliyon. 2023.e22531.
- [5] T. Kushartono, N. Fujilestari, and T. Muhammad Ali Ridho, "Political Journey of Indonesia: Dynamics and Challenges," *International Journal of Science and Research (IJSR)*, vol. 12, no. 12, pp. 734–740, Dec. 2023, doi: 10.21275/sr231207140923.
- [6] M. Adnan dan N. Mona, "Strategi Komunikasi Politik melalui Media Sosial oleh Calon Presiden Indonesia 2024," *Politika: Jurnal Ilmu Politik*, vol. 15, no. 1, pp. 1–20, Apr. 2024, doi: 10.14710/politika.15.1.2024.1-20.
- [7] Sekretariat Kabinet Republik Indonesia, "Hormati Jasa Para Pendahulu, Presiden Prabowo Anugerahkan Gelar Pahlawan Nasional kepada Sepuluh Tokoh Bangsa." Accessed: Jul. 07, 2026. [Online]. Available: <https://setkab.go.id/gelar-pahlawan-nasional-2025-prabowo/>
- [8] T. Putri, Siti Nurhaliza, and Dodi Vionanda, "Analisis Sentimen Penggunaan Aplikasi YouTube Menggunakan Metode Naïve Bayes," *UNP Journal of Statistics and Data Science*, vol. 3, no. 1, pp. 60–66, Feb. 2025, doi: 10.24036/ujsds/vol3-iss1/343.
- [9] A. S. Jessica Adinda Student, C. Author, and A. S. Jessica Adinda, "Journal of Humanities and Social Sciences Studies the Victim Narrative of the 1965-1966 Massacre in Indonesia as A Biopicture," 2022, doi: 10.32996/jhss.
- [10] N. Mahfudza and M. Ihksan, "Sentiment Analysis of YouTube Comments on Indonesian Presidential Candidates in 2024 using Naïve Bayes Classifier," *JURIKOM (Jurnal Riset Komputer)*, vol. 12, no. 2, pp. 140–148, Apr. 2025, doi: 10.30865/jurikom. v12i2.8538.
- [11] A. M. Firdaus, N. Ibrahim, and K. Kurniawati, "Hubungan Diplomatik Indonesia-Brunei Darussalam Pada Masa Presiden Suharto & Sultan Hassanah Bolkiah, 1984-1998," *Jurnal Humanitas: Katalisator Perubahan dan Inovator Pendidikan*, vol. 10, no. 1, pp. 140–155, Dec. 2023, doi: 10.29408/jhm. v10i1.23965.
- [12] R. Widjayanto and M. Huseini, "Transformasi Organisasi Sekretariat Kabinet Republik Indonesia dari Era Presiden Soeharto Sampai Presiden Joko Widodo," *Jurnal Manajemen dan Ilmu Administrasi Publik (JMIAP)*, vol. 5, no. 4, pp. 402–415, Dec. 2023, doi: 10.24036/jmiap. v5i4.777.
- [13] S. Ratnaswari, N. C. Wibowo, and D. S. Y. Kartika, "Analisis Sentimen Menggunakan Metode Lexicon-Based dan Support Vector Machine pada Presiden dan Wakil Presiden Indonesia Periode 2024–2029," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet. v13i1.5604.
- [14] J. E. Br Sinulingga and H. C. K. Sitorus, "Analisis Sentimen Opini Masyarakat terhadap Film Horor Indonesia Menggunakan Metode SVM dan TF-IDF," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 14, no. 1, pp. 42–53, Feb. 2024, doi: 10.34010/jamika.v14i1.11946.
- [15] S. Pesina and L. G. Yusupova, "Words Functioning in Lexicon," *Procedia Soc. Behav. Sci.*, vol. 192, pp. 38–43, Jun. 2015, doi: 10.1016/j.sbspro.2015.06.006.
- [16] I. Cero, J. Luo, and J. M. Falligant, "Lexicon-Based Sentiment Analysis in Behavioral Research," *Perspect. Behav. Sci.*, vol. 47, no. 1, pp. 283–310, Mar. 2024, doi: 10.1007/s40614-023-00394-x.

- [17] S. A. Rahmadhani, L. D. Rusanti, and H. Al Rosyid, "Klasifikasi Sentimen Komentar Youtube Demonstrasi DPR RI Menggunakan Support Vector Machine," *Arcitech: Journal of Computer Science and Artificial Intelligence*, vol. 5, no. 2, pp. 356–375, Dec. 2025, doi: 10.29240/arcitech.v5i2.15316.
- [18] P. C. Shilpa, R. Shereen, S. Jacob, and P. Vinod, "Sentiment analysis using deep learning," in *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 930–937. doi: 10.1109/ICICV50876.2021.9388382.
- [19] A. Novanto, D. Indra, and W. Astuti, "Analisis Pre-processing Sentimen Terhadap Komentar Layanan Indihome Pada Twitter," *Buletin Sistem Informasi dan Teknologi Islam*, vol. 5, no. 1, pp. 30–36, Feb. 2024, doi: 10.33096/busiti.v5i1.2066.
- [20] M. L. Lewis and M. C. Frank, "The length of words reflects their conceptual complexity," *Cognition*, vol. 153, pp. 182–195, Aug. 2016, doi: 10.1016/j.cognition.2016.04.003.
- [21] J. Pardede and D. Darmawan, "PERBANDINGAN ALGORITMA STEMMING PORTER, SASTRAWI, IDRIS, DAN ARIFIN & SETIONO PADA DOKUMEN TEKS BAHASA INDONESIA," vol. 12, no. 1, pp. 69–76, 2025, doi: 10.25126/jtiik.2025128860.
- [22] V. P. Carolina, E. Utami, and A. Yaqin, "Riset Jurnal Literatur: Morfologi Stemming Enhance Confix Stripping, Sastrawi Dan Tala Pada Bahasa Daerah Melayu-Ambon," *Journal homepage: Journal of Electrical Engineering and Computer (JEECOM)*, vol. xx, No. xx, pp. 186–194, Apr. 2024, doi: 10.33650/jeeecom.v4i2.
- [23] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian," *Electronics (Switzerland)*, vol. 11, no. 3, Feb. 2022, doi: 10.3390/electronics11030374.
- [24] Agung Teguh Wibowo Amais and Adi Susilo, "Principal Component Analysis-Based Data Clustering for Labeling of Level Damage Sector in Post-Natural Disasters," vol. XX, 2017, doi: 10.1109/ACCESS.2023.3275852.
- [25] A. Afda, "Analisis Dokumen Menggunakan Metode TF-IDF," *Jurnal Ekonomi Manajemen dan Informasi (JEMSI)*, vol. 5, pp. 466–169, Mar. 2024, doi: 10.31933/jemsi.v5i5.
- [26] S. M and Viera, "Cohen's Kappa Coefficient as a Performance Measure for Feature Selection," *International Conference on Fuzzy Systems (FUZZ-IEEE 2010)*, Jul. 2010, doi: 978-1-4244-8126-2.
- [27] C. A. Nurhaliza Agustina, R. Novita, Mustakim, and N. E. Rozanda, "The Implementation of TF-IDF and Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 156–163. doi: 10.1016/j.procs.2024.02.162.
- [28] P. Guleria, J. Frnda, and P. N. Srinivasu, "NLP-based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis," *Array*, vol. 27, Sep. 2025, doi: 10.1016/j.array.2025.100467.
- [29] J. R. Landis and G. G. Koch, "An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers," Jun. 1977. doi: <https://doi.org/10.2307/2529786>.
- [30] R. Zwick, "Another Look at Interrater Agreement," *Psychol. Bull.*, vol. 103, no. 3, pp. 374–378, 1988, doi: 10.1037/0033-2909.103.3.374.
- [31] D. Banerjee, T. A. Taffa, and R. Usbeck, "Best Practices in AI and Data Science Models Evaluation," in *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft für Informatik (GI)*, Gesellschaft für Informatik (GI), 2025, pp. 1211–1219. doi: 10.18420/inf2025_105.
- [32] M. Tella, M. Ahmed Usman, and K. Ibrahim Musa, "A Systematic Analysis of Performance Evaluation Metrics in Machine Learning Models," *INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)*, vol. XI, pp. 834–840, Jan. 2026, doi: 10.51584/IJRIAS.