

# **The Influence of Presidential Debate Comment Sentiment on YouTube on Candidate Electability: Naïve Bayes and Pearson Analysis**

**Arnoldus Yitzhak Petra Manoppo<sup>1</sup>, Wirawan Istiono<sup>2,\*</sup>**

<sup>1,2</sup>Informatics Department, Universitas Multimedia Nusantara, Tangerang, Indonesia  
Email: <sup>1</sup>arnoldus.manoppo@student.umn.ac.id, <sup>2</sup>wirawan.istiono@umn.ac.id

## **Abstract**

Campaigns significantly influence candidate electability. Presidential debates, a key campaign strategy, generate extensive public comments on social media, reflecting voter sentiment. This study employs VADER for automated sentiment labeling and Naïve Bayes for classification, analyzing comments from the KPU and Najwa Shihab YouTube channels. Electability data were sourced from national survey reports for correlation analysis. Pearson correlation results indicate that positive sentiment has a moderate negative correlation with electability, while negative sentiment shows a strong positive correlation. This suggests that negative sentiment in YouTube comments is more indicative of a candidate's rising electability, whereas positive sentiment does not necessarily translate into increased support. The Naïve Bayes model achieved 65% accuracy, 59% precision, 57% recall, and 57% F1-score when including neutral comments. Excluding neutral comments improved accuracy to 77%, with 68% precision, 68% recall, and 67% F1-score. The dataset comprised 17,872 comments, ensuring a robust sample. Despite these findings, limitations exist, such as potential biases in sentiment classification and representativeness, as social media users may not fully reflect the general voting population. Furthermore, while correlation is established, causality remains uncertain, requiring further research. This study enhances the understanding of social media sentiment in political campaigns and highlights the importance of integrating online sentiment analysis with traditional polling methods for a comprehensive assessment of electability.

**Keywords:** Machine Learning; Naïve Bayes; Pearson Correlation; Presidential Debate; Sentiment Analysis, Social Media Youtube

## **1. INTRODUCTION**

Elections represent a cornerstone of democracy, empowering citizens to shape their nation's leadership. In Indonesia, the 2024 presidential and legislative elections have become a high-stakes political contest marked by intensive campaigns and presidential debates. These debates, broadcast widely through channels like YouTube, provide a key stage for candidates to articulate their visions and respond to pressing national issues. Importantly, these online broadcasts attract thousands of comments from viewers, offering a real-time pulse of public sentiment. However, despite their abundance, these sentiments are often left



unanalyzed due to the sheer volume of data, leaving a significant gap in understanding the actual influence of debates on candidate electability [1].

The problem lies in the limited capacity to manually process and interpret vast amounts of public feedback on social media. YouTube, Indonesia's most used platform with 139 million users [2], becomes a primary arena for political discourse during elections. Channels such as the General Election Commission (KPU) and Najwa Shihab's official platform stream debate content, which in turn generates numerous viewer comments that reflect support, criticism, or neutrality toward the candidates. These expressions are not only reflections of viewer opinions but could serve as predictors of electoral outcomes. Nevertheless, despite this rich source of data, there is still a lack of comprehensive analysis linking these sentiments to actual candidate performance. This creates a clear research gap: while sentiments are visibly abundant, their impact on electability remains underexplored.

To bridge this gap, automated sentiment analysis using machine learning becomes a viable solution. Manually examining every comment is impractical, yet machine learning techniques such as Naive Bayes allow for rapid and scalable sentiment classification. Naive Bayes offers advantages like high efficiency on small datasets [3] and robustness with high-dimensional data [4],[5], making it well-suited for text-based sentiment analysis. Previous research has proven its efficacy in related contexts—hotel reviews [6], transportation sentiment [7], and even prior political sentiments on Twitter during the 2019 Indonesian presidential election [8]. Other related work, such as analyzing sentiment impact on stock prices using the KNN algorithm [9], has revealed patterns in how public opinion affects real-world outcomes.

Thus, this study aims to conduct sentiment analysis on YouTube comments related to Indonesia's presidential debates streamed by KPU and Najwa Shihab. Specifically, it investigates whether the sentiments expressed positive or negative correlate with candidate electability, and how these sentiments shape public perception of the debates. The research also evaluates the accuracy and reliability of the Naive Bayes model in this classification task. What sets this study apart is its integration of sentiment analysis with Pearson correlation, forming a dual approach to measure the relationship between online discourse and electoral traction.

Despite the promising approach, challenges remain. Sentiment analysis models may carry inherent biases and may not reflect the sentiments of the broader electorate. Furthermore, social media users represent only a subset of the population, limiting the generalizability of findings. Correlation-based results also stop short of establishing causation, highlighting the need for future longitudinal studies to determine how sentiments influence actual voting behavior.

Nevertheless, deriving insights from this study could guide more responsive campaign strategies and improve synergy between digital sentiment tracking and traditional political polling methods.

Sentiment analysis, a core application of Natural Language Processing (NLP), involves classifying text into positive, neutral, or negative categories [10]. Tools like Valence Aware Dictionary and Sentiment Reasoner (VADER) facilitate lexicon and rule-based sentiment analysis by evaluating the polarity of each word [11]. VADER outputs scores for positivity, neutrality, negativity, and a compound metric that normalizes overall sentiment [12], and is used here to label comment datasets automatically.

In parallel, data imbalance is addressed using the Synthetic Minority Oversampling Technique (SMOTE) [13]. This technique generates synthetic samples to balance underrepresented sentiment classes, ensuring the classifier does not become biased toward the majority class [14], [15]. Unbalanced data can cause significant misclassification errors [16], so balancing techniques like SMOTE are critical for robust model performance.

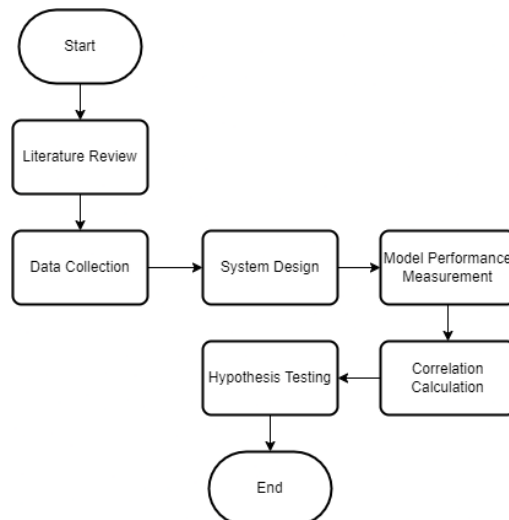
The Naive Bayes algorithm, rooted in Bayes' theorem, assumes feature independence and offers simplicity and efficiency in classification tasks [17], [18], [19]. To complement sentiment analysis, the Pearson Correlation Coefficient [20] is employed to quantify the linear relationship between sentiment polarity and electability, where values range from -1 (inverse correlation) to 1 (direct correlation), offering insights into the strength and direction of sentiment influence.

## 2. METHODS

The research methodology is illustrated in Figure 1, consisting of literature review, data collection, system design, model evaluation, correlation calculation, and hypothesis testing. To ensure clarity, the methodology diagram includes a representation of the literature review stage to emphasize its role in forming the theoretical basis for the study.

### 2.1. Literature Review

A comprehensive literature review was conducted to get gap of research and theories related to sentiment analysis, Naïve Bayes classification, and Pearson correlation in the context of electability studies.

**Figure 1.** Research Flow

## 2.2. Data Collection

The data collection process utilizes the YouTube Data API, as illustrated in Figure 2, to extract comments from the KPU and Najwa Shihab YouTube channels. The dataset consists of 17,872 comments, filtered based on candidate-related keywords and survey timelines. Additional preprocessing, such as filtering comments below 50 words, was applied to improve data relevance.

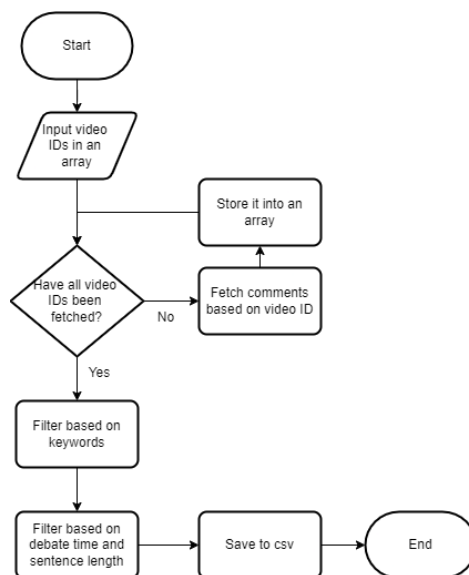
**Figure 2.** Flowchart of data collection

Figure 2 depicts the data collection flow. Data collection is performed on the YouTube comments column. Data collection is carried out using the YouTube Data API to retrieve comments on presidential debate videos based on video IDs. Comment data is filtered based on keywords for each candidate pair. Here are the keywords for each candidate pair: (1) Candidate pair 1: Anies, Imin, Amin. (2) Candidate pair 2: Prabowo, Gibran, Pragib. (3) Candidate pair 3: Ganjar, Mahfud, Gama. In addition to keywords, data is filtered based on the electability survey time, where only comments updated before the survey deadline are taken, and based on the word count below 50 words. Filtering based on word count is done to maximize model performance and remove comments that do not express opinions on the presidential debate results. The comment data is saved into a CSV file. In total, 17,872 comments were collected, divided into 15 datasets. Table 2 shows the videos used in data collection and the amount of data obtained.

**Table 2.** Data collection source

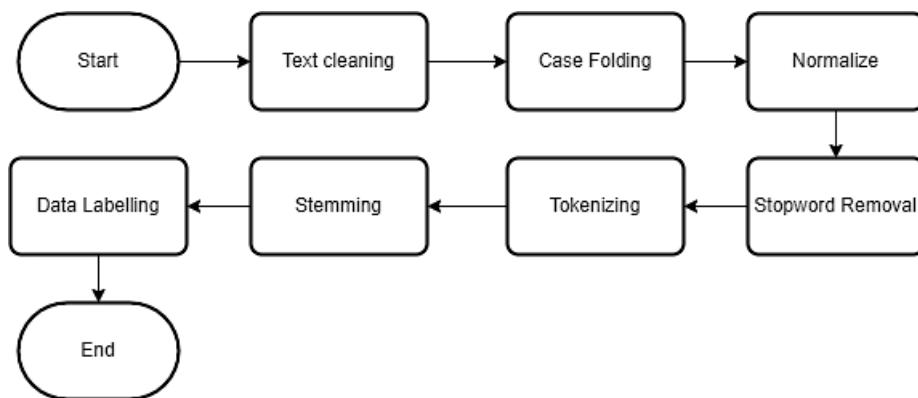
Debate	Channel	Title	Data
Presidential Candidate Debate 1	KPU RI	Debat Pertama Calon Presiden Pemilu Tahun 2024	7141
	Najwa Shihab	[LIVE] Musyawarah Nobar Debat Pilpres 2024   Musyawarah	
Vice Presidential Candidate Debate 2	KPU RI	Debat Kedua Calon Wakil Presiden Pemilu Tahun 2024	3825
	Najwa Shihab	[LIVE] Part 2 Nobar Debat Cawapres 2024   Musyawarah	
Presidential Candidate Debate 3	KPU RI	Debat Ketiga Calon Presiden Pemilu Tahun 2024	2986
	Najwa Shihab	[FULL] Debat Capres 2024, Nobar Debat Ronde Ketiga di Musyawarah   Musyawarah	
Vice Presidential Candidate Debate 4	KPU RI	Debat Keempat Calon Wakil Presiden Pemilu Tahun 2024	1972
	Najwa Shihab	[FULL] Debat Cawapres 2024, Nobar Debat Ronde Keempat di Musyawarah	
Presidential Candidate Debate 5	KPU RI	Debat Kelima Calon Presiden Pemilu Tahun 2024	1948
	Najwa Shihab	[FULL] Layar Tancap Mata Najwa, Nobar Debat Capres Ronde Kelima   Mata Najwa	
Total Data			17872

### 2.3. System Design

The system design incorporates VADER for automated sentiment labeling and Naïve Bayes for classification, as depicted in Figure 3. An improved flowchart illustrates the key stages of sentiment classification, including text preprocessing, feature extraction, and model training.

### 2.4. Data Preprocessing

The data preprocessing flow is shown in Figure 3. Data preprocessing consists of text cleaning, case folding, normalization, stopwords removal, tokenization, stemming, and data labeling [21].



**Figure 3.** Flowchart of data preprocessing

Figure 3 depicts the data preprocessing flowchart. Data preprocessing is a stage to prepare raw data. The purpose of data preprocessing is to process unstructured raw data into more structured data. There are several processes in data preprocessing as follow.

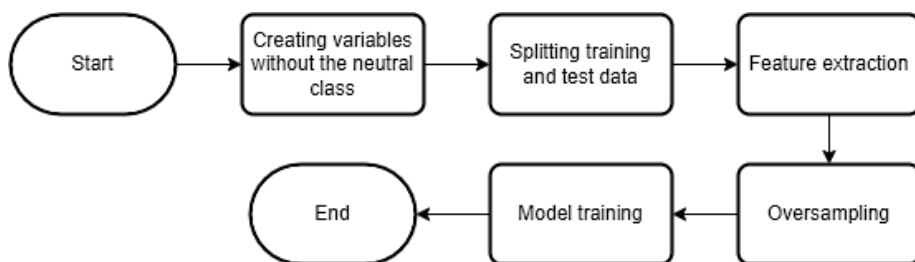
- 1) Text Cleaning: a process of text cleansing to reduce noise.
- 2) Case Folding: a process to standardize the case in a text, converting all characters to lowercase.
- 3) Normalize: a process to convert non-standard words into standard words.
- 4) Stopword Removal: a process to eliminate unimportant or meaningless words in a sentence.
- 5) Tokenizing: a process to segment text into tokens.
- 6) Stemming: a process to reduce a word to its base or root form.
- 7) Data Labeling: a process to assign sentiment labels to data. Sentiments are categorized into positive, neutral, and negative sentiments.

## 2.5. Naive Bayes Implementation

The implementation flow of Naive Bayes is shown in Figure 4. The implementation consists of creating variables without neutral class, splitting training and test data, feature extraction, oversampling, and model training. Equation 1 is the equation for Bayes' theorem.

$$P(C | X) = \frac{P(X | C) \cdot P(C)}{P(X)} \quad (1)$$

Where: X: data with an undefined class, C: hypothesis that data X belongs to a specific class,  $P(C|X)$ : posterior probability, or the probability of hypothesis C given data X,  $P(X|C)$ : likelihood, or the probability of X given the condition of hypothesis C, and  $P(X)$ : predictor prior probability, or the probability of X.



**Figure 4.** Flowchart of Naive Bayes implementation

Figure 4 depicts the Naive Bayes implementation flowchart. There are several stages in the implementation of Naive Bayes, namely:

- 1) Creating variables without neutral class: sentiment classification is conducted under two conditions: with neutral class and without neutral class.
- 2) Splitting training and test data: the data is divided into training and test sets. The training data comprises 80%, and the test data comprises 20% of the total data on dataset.
- 3) Feature extraction: after splitting the data into training and test sets, feature extraction is performed using Bag-of-Words (BOW).
- 4) Oversampling: after performing feature extraction, the next step is to perform oversampling on the minority class to address data imbalance.
- 5) Model training: After performing oversampling, the Naive Bayes model is trained. Naive Bayes then performs sentiment classification on the test data.

## 2.6. Model Evaluation

Model evaluation is conducted using accuracy, precision, recall, and F1-score metrics. A more detailed breakdown of these evaluation steps has been added to clarify how the Naïve Bayes model's performance is validated.

## 2.7. Correlation Calculation

The Pearson correlation coefficient is computed to measure the relationship between sentiment and electability, with an improved visualization to represent the statistical analysis process. The Pearson correlation value is obtained through the Equation 1 and guidelines for the degree of the relationship [22], as shown in Table 1.

$$r_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (2)$$

Where:

- x: the first variable
- y: the second variable
- n: the number of observations

**Table 1.** Guidelines for the Degree of Pearson Correlation

Correlation Coefficient	Degree of Relationship
0,00-0,199	Very Weak
0,20-0,399	Weak
0,40-0,599	Moderate
0,60-0,799	Strong
0,80-1	Very Strong

## 2.8. Hypothesis Testing

Hypothesis testing determines whether a significant relationship exists between social media sentiment and candidate electability. Clear criteria for accepting or rejecting the null hypothesis have been reinforced with explicit conditions for t-value comparisons. These hypotheses include:

- 1) H0: There is no relationship between presidential debate sentiment on social media and candidate electability.
- 2) H1: There is a relationship between presidential debate sentiment on social media and candidate electability.

The testing criteria with a significance level of 0.05 and degrees of freedom (n-2) are as follows.

- 1) If the t-value is positive:



- 2) (a) If  $t\text{-value} > \text{critical } t\text{-value (t-table)}$ , reject  $H_0$ .
- 3) (b) If  $t\text{-value} < \text{critical } t\text{-value (t-table)}$ , accept  $H_0$ .
- 4) If the  $t\text{-value}$  is negative:
- 5) (a) If  $-t\text{-value} < -\text{critical } t\text{-value (t-table)}$ , reject  $H_0$ .
- 6) (b) If  $-t\text{-value} > -\text{critical } t\text{-value (t-table)}$ , accept  $H_0$ .

These criteria determine whether to reject or accept the null hypothesis ( $H_0$ ) based on the calculated  $t\text{-value}$  compared to the critical  $t\text{-value}$  from the  $t\text{-table}$ .

### 3. RESULTS AND DISCUSSION

#### 3.1. Data Preprocessing

The data preprocessing stage is a crucial step to ensure the quality and reliability of sentiment analysis. This process includes a sequence of operations such as text cleaning, case folding, normalization, stopword removal, tokenization, stemming, translation, and data labeling. Initially, the dataset contained 17,872 comment entries. After applying all preprocessing techniques, the total number of valid entries was reduced to 16,884. To enhance readability and understanding, a consolidated table was created by merging the contents of Tables 4 through 10, with an additional column providing English translations of the Indonesian text. The transformation steps are detailed in the following tables. Text cleaning involves removing unnecessary punctuation such as commas and periods, which can interfere with the text analysis process. As shown in Table 3, the original sentence: “*Saya melihat sosok Anies Baswedan adalah sosok pemimpin yang mampu membawa Indonesia ke depan lebih baik ,Krn sejak menjadi gubernur DKI Jakarta , Jakarta menjadi lebih. Baik*” was cleaned to remove commas and periods, resulting in a smoother and cleaner version for processing.

**Table 3.** Text cleaning (Indonesia)

Before	After
Saya melihat sosok Anies Baswedan adalah sosok pemimpin yang mampu membawa Indonesia ke depan lebih baik ,Krn sejak menjadi gubernur DKI Jakarta , Jakarta menjadi lebih. Baik	Saya melihat sosok Anies Baswedan adalah sosok pemimpin yang mampu membawa Indonesia ke depan lebih baik Krn sejak menjadi gubernur DKI Jakarta Jakarta menjadi lebih Baik

Following text cleaning, case folding was performed. This step standardizes the text by converting all characters to lowercase, ensuring uniformity and reducing complexity for the model. Table 4 illustrates this transformation, where all capitalized words were converted to lowercase.

**Table 4.** Case folding (Indonesia)

Before	After
Saya melihat sosok Anies Baswedan adalah sosok pemimpin yang mampu membawa Indonesia ke depan lebih baik Krn sejak menjadi gubernur DKI Jakarta Jakarta menjadi lebih Baik	saya melihat sosok anies baswedan adalah sosok pemimpin yang mampu membawa indonesia ke depan lebih baik krn sejak menjadi gubernur dki jakarta jakarta menjadi lebih baik

Next, normalization was applied to correct informal or misspelled words into their proper forms. For instance, abbreviated words such as “krn” were normalized into “karena” (meaning “because”). Table 5 shows these corrections, which are essential to preserve the semantic integrity of the data.

**Table 5.** Normalize (Indonesia)

Before	After
saya melihat sosok anies baswedan adalah sosok pemimpin yang mampu membawa indonesia ke depan lebih baik krn sejak menjadi gubernur dki jakarta jakarta menjadi lebih baik	saya melihat sosok anies baswedan adalah sosok pemimpin yang mampu membawa indonesia ke depan lebih baik karena sejak menjadi gubernur dki jakarta jakarta menjadi lebih baik

The process continued with stopword removal, which aims to eliminate common words that do not contribute significantly to the sentiment of the sentence. Words such as “saya,” “adalah,” “yang,” and “karena” were removed, as seen in Table 6. This reduces noise and helps the classifier focus on more informative tokens.

**Table 6.** Stopword removal (Indonesia)

Before	After
saya melihat sosok anies baswedan adalah sosok pemimpin yang mampu membawa indonesia ke depan lebih baik karena sejak menjadi gubernur dki jakarta jakarta menjadi lebih baik	melihat sosok anies baswedan sosok pemimpin mampu membawa depan lebih baik sejak menjadi gubernur dki menjadi lebih baik

Following that, tokenization was conducted to break down each sentence into a list of individual words or tokens. This process, displayed in Table 7, transforms the sentence into an array of words, making it easier to analyze word frequency and structure. The next step, stemming, reduces words to their root forms to minimize redundancy. As shown in Table 8, terms like “melihat” become “lihat,” and “pemimpin” becomes “pimpin.” This step significantly decreases the dimensionality of the dataset, allowing for more efficient analysis.

**Table 7.** Tokenizing (Indonesia)

Before	After
melihat sosok anies baswedan sosok pemimpin mampu membawa depan lebih baik sejak menjadi gubernur dki menjadi lebih baik	['melihat', 'sosok', 'anies', 'baswedan', 'sosok', 'pemimpin', 'mampu', 'membawa', 'depan', 'lebih', 'baik', 'sejak', 'menjadi', 'gubernur', 'dki', 'menjadi', 'lebih', 'baik']

**Table 8.** Stemming (Indonesia)

Before	After
['melihat', 'sosok', 'anies', 'baswedan', 'sosok', 'pemimpin', 'mampu', 'membawa', 'depan', 'lebih', 'baik', 'sejak', 'menjadi', 'gubernur', 'dki', 'menjadi', 'lebih', 'baik']	lihat sosok anies baswedan sosok pimpin mampu bawa depan lebih baik sejak jadi gubernur dki jadi lebih baik

Once the preprocessing in Indonesian was complete, the text needed to be translated into English to be compatible with the VADER sentiment analysis tool. Table 9 shows the results of the translation process. For example, the sentence *“lihat sosok anies baswedan sosok pimpin mampu bawa depan lebih baik sejak jadi gubernur dki jadi lebih baik”* was translated to *“See the figure of Anies Baswedan, the figure of the leader is able to bring a better front since becoming the Governor of DKI to be better.”*

**Table 9.** Translating (Indonesia to English)

Before	After
lihat sosok anies baswedan sosok pimpin mampu bawa depan lebih baik sejak jadi gubernur dki jadi lebih baik	See the figure of Anies Baswedan, the figure of the leader is able to bring a better front since becoming the Governor of DKI to be better

**Table 10.** Data labelling

Text	Label
See the figure of Anies Baswedan, the figure of the leader is able to bring a better front since becoming the Governor of DKI to be better	Positive

Finally, data labeling was performed using the VADER sentiment analysis library. This step assigned sentiment categories—positive, neutral, or negative—to each translated comment. As shown in Table 10, the example text was labeled as “Positive.” This automated process ensures that each entry is sentiment-tagged based on the tone and polarity of the content, facilitating accurate sentiment classification.

### 3.2. Model Performance on Datasets with neutral Class

The Naïve Bayes model was trained using three different train-test split scenarios: 70:30, 80:20, and 90:10. The inclusion of multiple scenarios allows for a more comprehensive performance assessment. The model evaluation includes accuracy, precision, recall, and F1-score. Equation numbers have been added throughout the section for reference. Here are the classification report results for the dataset with the neutral class for candidate pair 1 in the first debate.

	precision	recall	f1-score	support
negative	0.32	0.50	0.41	60
neutral	0.82	0.58	0.68	166
positive	0.85	0.84	0.84	354
accuracy			0.74	580
macro avg	0.66	0.64	0.65	580
weighted avg	0.78	0.74	0.75	580

**Figure 5.** Classification report for dataset with neutral class

Figure 5 shows the classification report displaying overall accuracy and precision, recall, and F1 score for each class in the dataset with the neutral class. The results are as follows: overall accuracy is 74%, precision for the negative class is 32%, neutral class is 82%, positive class is 85%; recall for the negative class is 58%, neutral class is 58%, positive class is 84%; and F1 score for the negative class is 41%, neutral class is 68%, and positive class is 84%. Table 11 shows the overall model evaluation results for each candidate pair under the condition of neutral classes that still exist.

**Table 11.** Model evaluation for datasets with neutral class

Candidate	Accusation	Precision	Recall	F1-Score
Candidate 1	0,69	0,61	0,58	0,58
Candidate 2	0,62	0,61	0,58	0,58
Candidate 3	0,64	0,57	0,54	0,55
Average	65%	59%	57%	57%

Table 11 shows the model evaluation results for each candidate pair obtained from 15 datasets. An average accuracy of 65%, precision of 59%, recall of 57%, and F1 score of 57% were achieved. Table 12 shows the percentage of sentiment based

on 15 datasets for each candidate pair. The highest accuracy was obtained by candidate pair 1.

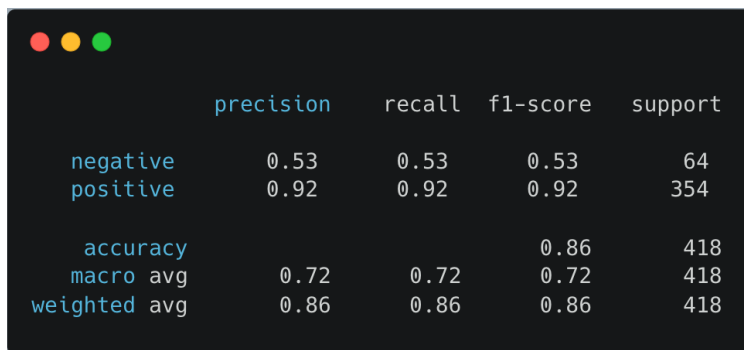
**Table 12.** Sentiment percentage for datasets with neutral class

Candidate	Positive	Neutral	Negative
Candidate 1	60,68	24,63	14,8
Candidate 2	50,06	21,04	28,8
Candidate 3	59,28	24,12	16,6

Table 12 shows the evaluation results for each candidate pair, indicating the average percentages of positive, neutral, and negative sentiments on the dataset with neutral classes. The highest positive sentiment was obtained by candidate pair 1, the highest neutral sentiment was obtained by candidate pair 1, and the highest negative sentiment was obtained by candidate pair 2.

### 3.3. Model Performance on Datasets without neutral Class

Here are the classification report results for the dataset without the neutral class for candidate pair 1 in the first debate.



	precision	recall	f1-score	support
negative	0.53	0.53	0.53	64
positive	0.92	0.92	0.92	354
accuracy			0.86	418
macro avg	0.72	0.72	0.72	418
weighted avg	0.86	0.86	0.86	418

**Figure 6.** Classification report for dataset without neutral class

Figure 6 shows the classification report displaying overall accuracy and precision, recall, and F1 score for each class in the dataset with the neutral class. The results are as follows: overall accuracy is 86%, precision for the negative class is 53%, positive class is 92%; recall for the negative class is 53%, positive class is 92%; and F1 score for the negative class is 53%, positive class is 92%. Table 13 shows the overall model evaluation results for each candidate pair with the neutral class removed.

**Table 13.** Model evaluation for datasets without neutral class

Candidate	Accuracy	Precision	Recall	F1-Score
Candidate 1	0,8	0,70	0,66	0,67
Candidate 2	0,73	0,67	0,68	0,67
Candidate 3	0,78	0,69	0,69	0,68
Average	77%	68%	68%	67%

Table 13 shows the model evaluation results for each candidate pair obtained from 15 datasets. An average accuracy of 77%, precision of 68%, recall of 68%, and F1 score of 67% were achieved. Based on Table 5, classification on the dataset with the neutral class removed, or binary classification, shows better results compared to classification on the dataset where the neutral class still exists, or multiclass classification. Table 14 shows the percentage of sentiment based on 15 datasets for each candidate pair. The highest accuracy was obtained by candidate 1.

**Table 14.** Sentiment percentage for datasets without neutral class

Candidate	Positive	Negative
Candidate 1	82,42	17,58
Candidate 2	65,4	34,6
Candidate 3	76,17	23,83

Table 14 shows the evaluation results for each candidate pair, indicating the average percentages of positive and negative sentiments on the dataset without the neutral class. The highest positive sentiment was obtained by candidate pair 1, and the highest negative sentiment was obtained by candidate pair 2.

### 3.4. Correlation Calculation and Hypothesis Testing

Sentiment analysis results are compared against actual election surveys. The Pearson correlation coefficient was computed to measure the relationship between sentiment and electability. The correlation calculation process is now elaborated with clearer equations and explanation. Table 15 shows variables X and Y for conducting correlation calculations obtained from the results of positive and negative sentiment percentages in the dataset with neutral classes.

**Table 15.** X and Y Variables

X positive	X negative	Y	Survey Agency
60,52%	19,00%	26,10%	CSIS
37,70%	41,00%	43,70%	
63,12%	16,00%	19,40%	
60,32%	15,00%	21,00%	Indicator

X positive	X negative	Y	Survey Agency
63,23%	14,00%	46,70%	IPS
59,55%	19,00%	24,50%	
59,22%	20,00%	21,30%	
52,61%	29,00%	51,80%	
63,91%	14,00%	19,20%	
66,36%	16,00%	21,30%	PWS
43,62%	36,00%	52,30%	
61,05%	18,00%	19,70%	
56,98%	4,00%	24,90%	Final Results
53,15%	24,00%	58,60%	
48,75%	16,00%	16,50%	

Table 15 shows variables X and Y along with the survey agency where:

- 1) X positive: percentage of positive sentiment
- 2) X negative: percentage of negative sentiment
- 3) Y: percentage of electability

Based on the values of X and Y shown in Table 15, the calculations are as follows.

- 1) n: 15
- 2)  $\sum XY$ : 25579.028
- 3)  $\sum X$ : 850.09
- 4)  $\sum Y$ : 467
- 5)  $\sum X^2$ : 49091.8031
- 6)  $\sum Y^2$ : 17598.26

$$r_{XY} = \frac{15\sum 25579,028 - (\sum 850,09)(\sum 467)}{\sqrt{[15\sum 49091,8031 - (\sum 850,09)^2][15\sum 17598,26 - (\sum 467)^2]}}$$

$$r_{XY} = \frac{-13306,61}{25094,34457} = -0,53$$

Based on the Pearson correlation coefficient guidelines shown in Table 1, the value of r indicates a moderate negative relationship. To determine if variable X is statistically significantly associated with variable Y, a t-test is conducted as follows.

$$t_{\text{count}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,53\sqrt{15-2}}{\sqrt{1-(-0,53)^2}} = -2,25$$

Because the calculated t-value (-2.25) is smaller than the critical t-value (-2.160) for (15-2) degrees of freedom, we reject the null hypothesis ( $H_0$ ), indicating that the relationship between variable X and variable Y is statistically significant. Based on the values of X and Y shown in Table 15, the calculations are as follows.

1. n: 15
2.  $\sum XY$ : 10577.5
3.  $\sum X$ : 301
4.  $\sum Y$ : 467
5.  $\sum X^2$ : 7241
6.  $\sum Y^2$ : 17598.26

$$r_{XY} = \frac{15\sum 10577,5 - (\sum 301)(\sum 467)}{\sqrt{[15\sum 7241 - (\sum 301)^2][15\sum 17598,26 - (\sum 467)^2]}}$$

$$r_{XY} = \frac{18095,5}{28750,14067} = 0,63$$

Based on the Pearson correlation coefficient guidelines shown in Table 1, the value of  $r$  indicates a strong positive relationship. To determine if variable X is statistically significantly associated with variable Y, a t-test is conducted as follows.

$$t_{\text{count}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,63\sqrt{15-2}}{\sqrt{1-0,63^2}} = 2,92$$

Because the calculated t-value (2.92) is greater than the critical t-value (2.160) for (15-2) degrees of freedom, we reject the null hypothesis ( $H_0$ ), indicating that the relationship between variable X and variable Y is statistically significant.

### 3.5. Discussion

The experimental results reveal several key insights into the application of sentiment analysis using the Naïve Bayes classifier on YouTube comments related to Indonesia's 2024 presidential debates. The process began with a comprehensive data preprocessing pipeline, which ensured the quality and relevance of the textual data. Through a series of transformations text cleaning, normalization, translation, and labeling raw comment data was converted into a machine-readable format suitable for sentiment classification. This foundational work supported the subsequent stages of model training and evaluation.

The performance of the Naïve Bayes classifier varied depending on the structure of the dataset, particularly in the inclusion or exclusion of the neutral sentiment class. On datasets where the neutral class was retained, the model achieved an average accuracy of 65% across all candidate pairs. While positive sentiment was



generally well-classified, the model struggled with the negative and neutral classes, as evidenced by lower precision and recall values. For example, the classification report for Candidate 1 (Figure 5) indicated that the F1 score for negative sentiment was only 0.41, compared to 0.84 for positive sentiment. This suggests that the model had difficulty distinguishing between less polarized opinions, which may be due to the linguistic ambiguity or overlapping sentiment expressions in comments labeled as neutral.

Conversely, performance improved notably on datasets without the neutral class. The model achieved a higher average accuracy of 77%, with consistent gains in precision, recall, and F1-score across all candidate pairs. For instance, Candidate 1 reached an F1-score of 0.67, a significant improvement compared to its performance on the multiclass dataset. This result aligns with expectations, as binary classification typically reduces complexity by removing the need to differentiate subtle or ambiguous sentiments. Table 13 further confirms this, showing that binary sentiment classification outperforms multiclass classification in all core metrics. The exclusion of the neutral class not only streamlined the learning process but also enhanced the model's sensitivity to positive and negative sentiment polarity.

Further analysis of sentiment percentages revealed meaningful distinctions among the candidate pairs. In both the multiclass and binary classification contexts, Candidate 1 consistently received the highest proportion of positive sentiment, with 60.68% in the multiclass dataset and 82.42% in the binary dataset. Meanwhile, Candidate 2 had the highest percentage of negative sentiment, suggesting a possible correlation between sentiment trends and public favorability. These observations provided the basis for a deeper investigation into the relationship between sentiment and electability.

To evaluate this connection, Pearson correlation analysis was conducted using sentiment percentages as variable X and electability survey results as variable Y. When using data from the multiclass dataset, the correlation coefficient  $r = -0.53$  indicated a moderate negative relationship between negative sentiment and electability. The corresponding t-test showed that this correlation was statistically significant, suggesting that higher negative sentiment is associated with lower electability outcomes. This result supports the hypothesis that online discourse reflects real-world political dynamics.

On the other hand, the correlation using positive sentiment as X yielded a coefficient of  $r = 0.63$ , indicating a strong positive relationship. The calculated t-value of 2.92 exceeded the critical value, confirming statistical significance. This finding reinforces the idea that positive sentiment in social media comments is strongly aligned with higher electability ratings in survey results. It implies that

sentiment analysis when supported by reliable preprocessing and validated with correlation testing can be a powerful proxy for measuring public opinion during elections.

Taken together, the results suggest that binary classification models provide higher accuracy and better interpretability for sentiment analysis in political contexts. The inclusion of neutral sentiment adds nuance but introduces classification challenges that reduce model performance. Furthermore, the strong correlation between sentiment and electability supports the hypothesis that social media sentiment can serve as an indicator of public support, aligning with conventional survey findings. These insights open opportunities for further integrating sentiment analysis into political campaign strategy, public opinion monitoring, and election forecasting.

However, several limitations must be acknowledged. The reliance on translated comments may introduce semantic distortion, as not all Indonesian expressions translate clearly into English. Additionally, the sentiment analysis model, though effective, is limited to surface-level polarity detection and may not fully capture sarcasm, context, or deeper political nuance. Social media data is also inherently biased, as it represents only active online users and not the entire voting population. Lastly, while correlation was established, causation cannot be inferred, and future research is needed to explore causative links between sentiment and electoral outcomes using longitudinal data or hybrid methodologies. Overall, the study demonstrates that a carefully structured approach combining NLP preprocessing, machine learning classification, and statistical correlation can yield meaningful insights into the role of public sentiment in elections. These findings underscore the growing value of digital sentiment analysis as a complementary tool to traditional political science methods.

#### 4. CONCLUSION

This study analyzed the relationship between sentiment in YouTube comments on presidential debates and candidate electability using sentiment analysis and Pearson correlation. The findings reveal that positive sentiment has a moderate negative correlation with electability, whereas negative sentiment has a strong positive correlation, both of which are statistically significant. This suggests that negative sentiment may indicate higher public attention and engagement with candidates, potentially influencing electability. The sentiment distribution analysis showed that candidate pair 1 received the highest positive sentiment in both datasets with and without neutral classes, while candidate pair 2 had the highest negative sentiment. Model evaluation demonstrated that the Naïve Bayes algorithm achieved an average accuracy of 65% in datasets with neutral classes and 77% in datasets without neutral classes, indicating better classification performance in binary sentiment classification.

For political strategists, these findings highlight the importance of monitoring and analyzing negative sentiment, as it may be an indicator of public awareness and potential voter shifts. Future research could explore alternative classification models such as Deep Learning or Transformer-based architectures to improve sentiment analysis accuracy. Additionally, comparing sentiment trends across different social media platforms like Twitter, Facebook, and Instagram may provide a broader perspective on voter sentiment. Expanding the dataset and integrating demographic factors could further enhance the understanding of how online discussions shape political outcomes. While this study provides valuable insights, limitations such as potential biases in social media data and the lack of causal inference should be addressed in future research. A mixed-method approach combining sentiment analysis with qualitative voter surveys could further validate the influence of social media sentiment on candidate electability.

## REFERENCES

- [1] F. P. Bariguna, A. Sulaeman, and W. Budi Darmawan, "Electoral Behavior in The Electability of Presidential and Vice Presidential Candidates in The 2019 Elections," *JIP (Jurnal Ilmu Pemerintahan) Kaji. Ilmu Pemerintah. dan Polit. Drb.*, vol. 6, no. 1, pp. 13–22, 2021, doi: 10.24905/jip.6.1.2021.13-22.
- [2] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: Geographic popularity of videos," *Proc. 21st Annu. Conf. World Wide Web*, pp. 241–250, 2012, doi: 10.1145/2187836.2187870.
- [3] A. Meiriza, E. Lestari, P. Putra, A. Monaputri, and D. A. Lestari, "Prediction Graduate Student Use Naive Bayes Classifier," vol. 172, no. Siconian 2019, pp. 370–375, 2020, doi: 10.2991/aisr.k.200424.056.
- [4] S. J. Simoff, G. J. Williams, J. Galloway, and I. Kolyshkina, *A U S D M O 5 Edited by*, no. May. 2014.
- [5] M. Zhikri and W. Istiono, "Handling Class Imbalance for Indonesian Twitter Sentiment Analysis A Comparative Study of Algorithms," *J. Syst. Manag. Sci.*, vol. 14, no. 10, pp. 170–179, 2024, doi: 10.33168/JSMS.2024.1010.
- [6] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012024.
- [7] S. Afrizal, H. N. Irmanda, N. Falih, and I. N. Isnainiyah, "Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap Kehadiran Mass Rapid Transit," *J. Inform.*, vol. 15, no. 3, pp. 157–168, 2019.
- [8] S. N. J. Fitriyyah, N. Safriadi, and E. E. Pratama, "Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 279, 2019, doi: 10.26418/jp.v5i3.34368.

- [9] M. G. Pradana, A. C. Nurcahyo, and P. H. Saputro, "Pengaruh Sentimen Di Sosial Media Dengan Harga Saham Perusahaan," *J. Ilm. Edutic*, vol. 6, no. 2, 2020, doi: 10.21107/edutic.v6i2.6992.
- [10] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *J. SIMETRIS*, vol. 10, no. 2, pp. 681–686, 2019.
- [11] H. Sagala and H. Toba, "Penentuan Aspek yang Berpengaruh Terhadap Produk Smartphone Berdasarkan Ulasan Berbasis Teksstual," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 1, pp. 287–295, 2021, doi: 10.28932/jutisi.v7i1.3466.
- [12] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for," *Eighth Int. AAAI Conf. Weblogs Soc. Media*, pp. 216–225, 2014.
- [13] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.
- [14] N. N. Sholihah and A. Hermawan, "Implementation of Random Forest and Smote Methods for Economic Status Classification in Cirebon City," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1387–1397, 2023, doi: 10.52436/1.jutif.2023.4.6.1135.
- [15] R. Peranginangin, E. J. G. Harianja, I. K. Jaya, and B. Rumahorbo, "Penerapan Algoritma Safe-Level-Smote Untuk Peningkatan Nilai G-Mean Dalam Klasifikasi Data Tidak Seimbang," *METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt.*, vol. 4, no. 1, pp. 67–72, 2020, doi: 10.46880/jmika.vol4no1.pp67-72.
- [16] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00349-y.
- [17] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, pp. 787–795, 2013, doi: 10.2478/amcs-2013-0059.
- [18] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Comparison of Naïve Bayes and Support Vector Machine Methods in Twitter Sentiment Analysis," *Smatika J.*, vol. 10, no. 02, pp. 71–76, 2020.
- [19] M. N. Randhika, J. C. Young, A. Suryadibrata, and H. Mandala, "Implementasi Algoritma Complement dan Multinomial Naïve Bayes Classifier Pada Klasifikasi Kategori Berita Media Online," *Ultim. J. Tek. Inform.*, vol. 13, no. 1, pp. 19–25, 2021, doi: 10.31937/ti.v13i1.1921.
- [20] Miftahuddin, A. Pratama, and I. Setiawan, "Analisis Hubungan Antara Kelembaban Relatif Dengan Beberapa Variabel Iklim," *J. Siger Mat.*, vol. 02, no. 01, pp. 25–33, 2021.
- [21] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, pp. 776–784, 2022, doi: 10.11591/ijece.v12i1.pp776-784.

- [22] F. Jabnabillah and N. Margina, “Analisis Korelasi Pearson Dalam Menentukan Hubungan Antara Motivasi Belajar Dengan Kemandirian Belajar Pada Pembelajaran Daring,” *J. Sintak*, vol. 1, no. 1, pp. 14–18, 2022.