Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025 e-ISSN: **2656-4882** p-ISSN: **2656-5935**
DOI: 10.51519/journalisi.v7i1.1019 Published By DRPM-UBD

Evaluation of Machine Learning Models for Sentiment Analysis in the South Sumatra Governor Election Using Data Balancing Techniques

Febriyanti Panjaitan[1], Win Ce[2], Hery Oktafiandi[3], Ghanim Kanugrahan[4], Yudi Ramdhani[5], Vito Hafizh Cahaya Putra[6]

[1,4,6]Informatics Departement, Satu University, Bandung, Indonesia
[2]Information System Departement, Bina Nusantara University, Bandung, Indonesia
[3,5]Information System Departement, Satu University, Bandung, Indonesia
Email: [1*]pebrianti.panjaitan@univ.satu.ac.id

Abstract

Sentiment analysis is crucial for understanding public opinion, especially in political contexts like the 2024 South Sumatra gubernatorial election. Social media platforms such as Twitter and YouTube provide key sources of public sentiment, which can be analyzed using machine learning to classify opinions as positive, neutral, or negative. However, challenges such as data imbalance and selecting the right model to improve classification accuracy remain significant. This study compares five machine learning algorithms (SVM, Naïve Bayes, KNN, Decision Tree, and Random Forest) and examines the impact of data balancing on their performance. Data was collected via Twitter crawling (140 entries) and YouTube scraping (384 entries), and text features were extracted using CountVectorizer. The models were then evaluated on imbalanced and balanced datasets using accuracy, precision, recall, and F1-score. The Decision Tree and Random Forest models achieved the highest accuracies of 79.22% and 75.32% on imbalanced data, respectively. However, they also exhibited overfitting, as indicated by their near-perfect training performance. Naïve Bayes, on the other hand, demonstrated the lowest accuracy at 54.55% despite achieving high precision, suggesting frequent misclassification, particularly for the minority class. SVM and KNN also struggled with imbalanced data, recording accuracies of 58.44% and 63.64%, respectively. Significant improvements were observed after applying data balancing techniques. The accuracy of SVM increased to 71.43%, and KNN improved to 66.23%, indicating that these models are more stable and effective when class distributions are even. These findings highlight the substantial impact of data balancing on model performance, particularly for methods sensitive to class distribution. While tree-based models achieved high accuracy on imbalanced data, their tendency to overfit underscores the importance of balancing techniques to enhance model generalization.

Keywords: Sentiment Analysis, Machine Learning, Governor Election, Text Analysis, Crawling, Algorithm, Balancing Data.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

1. INTRODUCTION

The 2024 South Sumatra gubernatorial election is a pivotal moment in local democracy, where public opinion significantly influences the direction of regional leadership [1]. With the rapid advancement of digital technology and social media, the public increasingly expresses opinions regarding gubernatorial candidates, work programs, and political issues [2]. Social media platforms like Twitter and YouTube are primary channels for sharing political perspectives, criticisms, and support for competing candidates [3]. The large volume of information generated on these platforms presents an opportunity for sentiment analysis to assess voter preferences and track political trends [4] [5].

One of the main challenges in sentiment analysis is data imbalance, where the distribution of positive, negative, and neutral sentiments is uneven. This imbalance can reduce the effectiveness of Machine Learning models, leading to biased predictions and suboptimal classification results [6]. In sentiment analysis research, imbalanced datasets often cause models to favor dominant sentiment classes while underrepresenting minority classes, which can lead to misleading conclusions [7]. Previous studies have shown that addressing data imbalance through balancing techniques can significantly improve model performance in sentiment classification [8], [9].

Sentiment analysis using Machine Learning methods has become a widely used research approach in political and social fields [8]. Various Machine Learning algorithms, such as Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest, can be used to classify public sentiment toward gubernatorial candidates [10]. However, beyond data imbalance, other challenges include the high dimensionality of textual data, language use variations, and noise in social media text, which can impact model accuracy and generalization ability [11].

Previous studies have demonstrated the effectiveness of Machine Learning methods in sentiment analysis, particularly in politics and elections. SVM is known for its high performance in text classification and has been used in sentiment analysis research on elections, as shown in [12], where SVM achieved higher accuracy than other text-based sentiment classification methods. Naïve Bayes, as a probabilistic method, is also widely used in sentiment analysis due to its ability to handle large text datasets and its efficiency in processing, as evidenced by the study in [13] on sentiment analysis on Twitter. Moreover, the KNN algorithm has been widely applied in various sentiment analysis studies due to its simplicity and effective capability in classification based on data proximity. As demonstrated in [14], KNN provides competitive results in text classification. Decision Tree is also frequently used in sentiment analysis because of its high interpretability and ability

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

to handle datasets with complex features. This was discussed in [15], highlighting Decision Tree's effectiveness in classifying sentiments across various social media platforms. Meanwhile, as an ensemble-based method, Random Forest has been proven to be more stable and accurate compared to other single models in many studies [16], indicating that Random Forest performs better in handling sentiment variations in Twitter sentiment analysis.

Based on findings from previous studies, the SVM, Naïve Bayes, KNN, Decision Tree, and Random Forest methods were selected for this research because they have been proven to possess distinct advantages in text-based sentiment analysis. Although numerous studies have employed Machine Learning methods for political sentiment analysis, several challenges remain, particularly regarding data imbalance, which frequently occurs in public opinion datasets. Most previous studies have focused solely on algorithm performance without considering the impact of data distribution on model effectiveness. Therefore, this study offers novelty by exploring how these five Machine Learning methods (SVM, Naïve Bayes, KNN, Decision Tree, and Random Forest) perform under both balanced and imbalanced data conditions, specifically in the context of the 2024 South Sumatra gubernatorial election. Consequently, this research evaluates model effectiveness and provides insights into data imbalance handling strategies to enhance the accuracy of political sentiment analysis.

The urgency of this research lies in the increasing use of social media as the primary platform for the public to express political opinions. Information disseminated on Twitter and YouTube can reflect voter preferences and emerging political trends, making sentiment analysis based on Machine Learning a crucial tool for understanding real-time election dynamics. This study aims to develop a more accurate and reliable sentiment analysis model, which will benefit academics, researchers, policymakers, campaign teams, and the media in better understanding and responding to public opinion.

2. METHODS

This research utilizes a Machine Learning approach to analyze public sentiment toward the 2024 South Sumatra Gubernatorial Election based on opinions collected from social media. Data was obtained through Twitter crawling and YouTube comment scraping, then processed through several stages to ensure data quality and relevance before being applied to the Machine Learning model.

As illustrated in Figure 1, the research methodology consists of several key stages: data collection, preprocessing, feature extraction, handling imbalanced data with RandomOverSampling, Machine Learning modeling, and Evaluation. The data collection involves retrieving relevant tweets and YouTube comments, which are

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

then cleaned and processed to remove noise. The preprocessing stage includes tokenization, stopword removal, and stemming to improve text quality. Feature extraction uses CountVectorizer to transform textual data into numerical representations, capturing single words and frequent word combinations (bigrams and trigrams). This approach has been widely used in sentiment analysis research [17], [18], [19] to enhance contextual understanding and improve model accuracy.

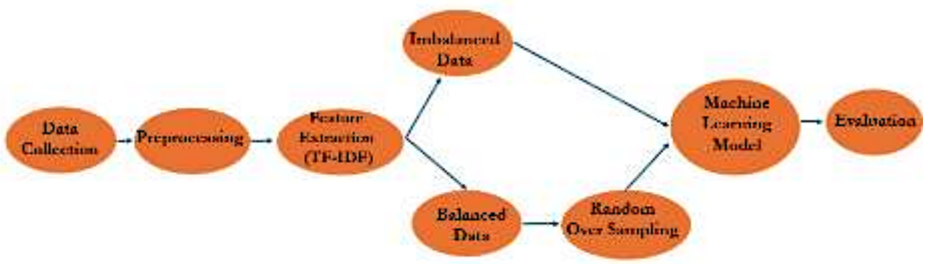

Figure 1. Methodology

Each stage is described in detail as follows.

2.1. Data Collection (Crawling dan Scraping)

In this study, the data used was obtained from two main sources:

1) Data was obtained using web crawling techniques via Twitter API or libraries like Tweepy. The crawling process aimed to gather tweets containing public opinions related to the 2024 South Sumatra Governor Election, covering the period from October 2024 to January 2025. To ensure relevance, a set of carefully selected keywords was used as filters, including:
   a) The names of gubernatorial candidates,
   b) Election-related hashtags (e.g., #PilkadaSumsel2024, #DebatGubernurSumsel),
   c) Frequently used phrases in political discussions related to the election.

   The collected tweets were then preprocessed to remove noise, such as duplicate entries, non-relevant content, and automated bot-generated tweets, ensuring that the dataset primarily contained organic public opinions.

2) YouTube: Comment data was retrieved using web scraping techniques through the YouTube API, complemented with Selenium to extract comments directly from selected videos discussing the 2024 South Sumatra Governor Election. The primary focus was on videos from the official *Komisi Pemilihan Umum* (KPU) YouTube account, particularly those covering the Governor Debates. These videos were chosen based on the following criteria:

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

a) They had at least 100 comments, ensuring a substantial volume of public discourse.
b) They were directly related to the election, such as candidate debates, political analyses, or candidate interviews.

The extracted comments underwent a preprocessing phase, including removing irrelevant content (e.g., spam, advertisements), tokenization, and language filtering to retain only relevant textual data reflecting public sentiment towards the gubernatorial candidates**.** By combining data from Twitter and YouTube, this study leverages diverse online discourse to provide a comprehensive sentiment analysis of public opinion in the 2024 South Sumatra Governor Election.

2.2. Preprocessing Data

Before data can be used for sentiment analysis, preprocessing must be performed to ensure that the text is cleaner and more structured. These preprocessing steps have been applied in several studies [20], [21] for sentiment analysis, resulting in higher-quality data that can be accurately labelled with the sentiment. Text preprocessing is a crucial step in sentiment analysis, ensuring that textual data is cleaned and structured for accurate classification by machine learning models. The process involves several key steps, each designed to refine and standardize the text before sentiment labeling.

The first step is case folding, which converts all text to lowercase. This helps prevent discrepancies between uppercase and lowercase letters, ensuring that words such as "Election" and "election" are treated as the same entity. By applying case folding, inconsistencies in capitalization are eliminated, making the data more uniform and easier to process. Next, removing unnecessary characters is performed to clean the text of elements that do not contribute to sentiment analysis. This step includes removing punctuation marks, numbers, symbols, emojis, and other special characters. Additionally, mentions (usernames), URLs, and hashtags are eliminated to focus solely on the textual content. By filtering out these elements, the dataset remains relevant and free from distractions that could affect sentiment classification.
T
okenization follows, where the text is split into individual words or tokens. This step is essential because machine learning models analyze sentiment at the word level, and breaking down sentences into smaller units enables more effective processing. Tokenization helps structure the text in a way that allows algorithms to recognize patterns and relationships between words. Once tokenization is complete, stopword removal is applied to eliminate common words that do not add significant meaning to sentiment analysis. Words such as "that," "and," "or," and "in" are removed using predefined stopword lists from tools like NLTK or

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Sastrawi. By excluding these words, the model focuses on more meaningful terms that influence sentiment classification.

Another important step is stemming, which reduces words to their root form. For example, the Indonesian word "*memilih*" (choosing) is simplified to "*pilih*" (choose). This process helps reduce data complexity and ensures that variations of the same word are treated as a single entity. The Sastrawi Stemmer algorithm is commonly used for stemming in the Indonesian language, making it a valuable tool for text normalization. Finally, after preprocessing, each text is assigned a sentiment label based on its content. A text is classified as positive if it expresses support or favorable opinions about the governor candidate. Conversely, a text is labeled as negative if it contains criticism or unfavorable opinions. If a text lacks clear sentiment or is purely informational, it is categorized as neutral. This labeling process enables sentiment analysis models to categorize opinions accurately, providing valuable insights into public perception. Sentiment labelling can be conducted using a lexicon-based approach with a sentiment dictionary to expedite the initial annotation process [22].

2.3. Imbalance and Balance Data

The effectiveness of sentiment analysis models can vary significantly depending on the distribution of the dataset. To comprehensively assess model performance, this study evaluates sentiment classification under two different data distribution scenarios: balanced data and imbalanced data.

1) Balanced Data → Sentiment data is balanced using methods such as RandomOverSampling to ensure that each class has the same data points.
   a) RandomOverSampling is a technique for handling data imbalance by increasing the number of samples from the minority class. This technique works by randomly duplicating examples from the underrepresented class until its proportion is balanced with that of the majority class. By doing so, the machine learning model is not biased toward the class with more samples.
   b) Previous studies [23], [24] have demonstrated that RandomOverSampling effectively enhances model performance in handling imbalanced sentiment datasets. This method is computationally efficient and easy to implement, making it a preferred choice for large-scale text datasets in sentiment analysis.
   c) However, RandomOverSampling carries the risk of overfitting since the newly added data are duplicates of existing ones. To mitigate this, it is often combined with other methods, such as SMOTE (Synthetic Minority Over-sampling Technique), which generates more diverse synthetic samples to enhance generalization capability [25].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2) Imbalanced Data → The model is tested under real-world conditions where the number of positive, negative, and neutral sentiment data is uneven [26].

By comparing the results from both scenarios, this study aims to identify the most effective algorithm for handling data imbalance and provide recommendations for optimizing political sentiment analysis using Machine Learning.

2.4. Feature Extraction

Once the data is ready for use, text features are converted into numerical forms using CountVectorizer to be processed by Machine Learning algorithms [27]. CountVectorizer counts the frequency of words or phrases in a document without considering their rarity weight, as in TF-IDF. The steps are as follows:
1) Tokenization – Splitting text into words or phrases based on the specified N-Gram.
2) Feature Vector Creation – Counting the occurrences of each word or word combination in the document to form a numerical representation.
3) Applying CountVectorizer – Using CountVectorizer from the Scikit-learn library to transform the text into a feature matrix that can be utilized in Machine Learning modelling.

2.5. Machine Learning Modeling

After completing the preprocessing and feature extraction stages using CountVectorizer, the data is ready for training the Machine Learning model. The model is trained using processed training data to recognize patterns in text and classify sentiment more accurately. CountVectorizer converts text into numerical vectors representing the weight of words in the document, allowing the model to differentiate words with significant influence in sentiment analysis [28]. After feature extraction, the modelling process is carried out using several Machine Learning algorithms, namely:
1) Support Vector Machine (SVM)
2) Naïve Bayes
3) K-Nearest Neighbors (KNN)
4) Decision Tree
5) Random Forest

2.6. Evaluation and Performance Comparison of the Model

To assess the effectiveness of the model, an evaluation is conducted based on the following metrics [29], [30]:
1) Accuracy → Measures how correctly the model classifies sentiment.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2) Precision, Recall, and F1-Score → Evaluate the balance between positive and negative classifications and the model's performance on imbalanced data.
3) Confusion Matrix → Displays the correct and incorrect predictions for each sentiment category.

The dataset is partitioned into 80% for training and 20% for testing. This 80:20 split is a widely accepted practice in machine learning research, as it provides sufficient data for model training while ensuring a robust and unbiased evaluation of the test set. Such a division has been demonstrated to produce reliable and accurate assessments of model performance in numerous studies [31], [32].

3. RESULTS AND DISCUSSION

This section presents the data collection process results, preprocessing, sentiment labelling, and Machine Learning model training for sentiment analysis in the 2024 South Sumatra Governor Election. The data used was obtained through Twitter crawling and YouTube comment scraping, then processed through several stages to be ready for modelling.

This study's findings include the amount of data obtained from both sources, the preprocessing steps performed to clean and optimize the data, the sentiment distribution found in the dataset, and the application of Machine Learning algorithms to analyze sentiment patterns. Additionally, a performance comparison of the models on balanced and imbalanced data was conducted to identify the most optimal model for handling political sentiment analysis.

3.1 Data Collection

In the data collection phase, crawling was conducted on Twitter and scraping on YouTube to gather public opinions regarding the 2024 South Sumatra Governor Election. Table 1 presents the number of data collected from Twitter and YouTube before and after the data cleaning process. From Twitter, 140 data points were obtained, and none were removed as all were valid. Meanwhile, from YouTube, 398 data points were initially collected, but 14 were removed due to containing NaN values (empty or unusable), resulting in a final count of 384 data points used in the study. NaN data was removed to ensure that only valid and relevant data were used in the sentiment analysis. Thus, the total number of data points used after cleaning was 524, further processed in the preprocessing and Machine Learning modelling stages. The details of the data count from each source are shown in Table 1.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

Table 1. Collected Data

No	Data Source	Total Data	Data After Removing NaN	Data Removed (NaN)
1	Twitter (Crawling)	140	140	0
2	YouTube (Scraping)	398	384	14
	Total	538	524	14

Data collection from these two platforms aims to obtain a richer perspective on public sentiment toward the gubernatorial candidates.

3.2 Preprocessing Data

Before conducting sentiment analysis, the collected data undergoes preprocessing to improve its quality. The preprocessing stages include case folding, tokenization, stopword removal, and stemming. Table 2 presents the results of the preprocessing stages applied to data collected from Twitter and YouTube. This process aims to clean and simplify the text to be processed more effectively in sentiment analysis using machine learning.

1) Raw Data: The raw data from Twitter still contains mentions (@username), capital letters, and special characters such as punctuation marks and emojis. Data from YouTube also includes emojis (🔥🔥🔥) that need to be removed as they do not contribute to text-based sentiment analysis.
2) Case Folding: All text is converted to lowercase to ensure consistency in text processing. For example, text from Twitter that originally contained capitalized words is entirely converted to lowercase.
3) Tokenization: Sentences are broken down into individual words for further analysis. At this stage, a list would split a sentence like "masih menunggu prediksi dari pengamat politik" into separate elements.
4) Stopword Removal: Common words that do not carry significant meaning in sentiment analysis, such as "still", "from", "we", "will", "in", are removed. This aims to eliminate words that do not provide essential sentiment-related information in the text.
5) Stemming: Words are reduced to their root forms using stemming techniques. For example, the word "kesayangan" becomes "Sayang," and "pengamat" becomes"amat," allowing the model to recognize root words more effectively. The "edi santana menyala" remains unchanged in the YouTube data, as no words require modification.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

Table 2. Example of Data Preprocessing Results

Stage	Original Text (Twitter)	Original Text (YouTube)
Raw Data	*Masih menunggu prediksi dari pengamat politik kesayangan kita semua ttg siapa yg akan menang di pilkada sumsel* @tmypnc @bagaskite @togisatrio @taufikags @halim_mhas	*Edi Santana menyala* 🔥🔥🔥
Case Folding	*masih menunggu prediksi dari pengamat politik kesayangan kita semua ttg siapa yg akan menang di pilkada sumsel*	*edi santana menyala*
Tokenization	[*'masih ', 'menunggu','prediksi', 'dari', 'pengamat', 'politik', 'kesayangan', 'kita', 'akan', 'ttg', 'siapa', 'yg', 'menang', 'di', 'pilkada', 'sumsel'*]	*['edi', ' santana ',' menyala']*
Stopword Removal	*['tunggu', 'prediksi', 'amat', 'politik', 'sayang', 'ttg', 'menang', 'pilkada', 'sumsel']*	*['edi', ' santana ',' menyala']*
Stemming	*tunggu prediksi amat politik sayang ttg menang pilkada sumsel*	*edi santana menyala*

3.3 Sentiment Labeling

After preprocessing, the data is assigned sentiment labels based on three main categories: neutral, positive, and negative. The labelling process is carried out automatically using the sentiment_analysis_lexicon_indonesia method, which utilizes a sentiment-oriented word dictionary in the Indonesian language to determine the polarity of each text.

As shown in Table 3, most of the data has a neutral sentiment, both on Twitter (70 out of 140 data points) and YouTube (264 out of 384 data points). This indicates that many opinions are informative or descriptive, without expressing strong positive or negative feelings toward the gubernatorial candidate. Additionally, positive sentiment data is more dominant than negative sentiment on Twitter (53 positive, 18 negative) and YouTube (63 positive, 30 negative). This suggests that the collected public opinions express more support or optimism toward the gubernatorial candidate rather than criticism or dissatisfaction.

Although lexicon-based analysis has advantages in speed and ease of implementation, this method also has limitations, such as difficulties in handling words with multiple meanings (ambiguity) and sarcasm. Therefore, the labelling results will be a foundation for Machine Learning modelling, which can improve sentiment prediction accuracy by understanding more complex contexts.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

Table 3. Sentiment Label Distribution

Data Source	Neutral	Positive	Negative	Total
Twitter (Crawling)	70	53	18	140
YouTube (Scraping)	232	109	43	384
Total	302	162	61	524

3.4 Evaluation and Performance Comparison of Models

Tables 4 and 5 present the evaluation results of Machine Learning models on imbalanced and balanced data. The evaluation is based on accuracy, precision, recall, and F1-score on test and training data for five Machine Learning methods: SVM, Naïve Bayes, KNN, Decision Tree, and Random Forest.

1) Model Performance on Imbalanced Data

Decision Tree and Random Forest demonstrated the best performance on imbalanced data, with test accuracy rates of 79.22% and 75.32%, respectively, along with high F1-scores (78.91% and 72.41%). This indicates that both models have strong classification capabilities despite the unequal class distribution in the dataset. Conversely, Naïve Bayes had the lowest test accuracy (54.55%), although it achieved a high precision score (77.41%). This suggests that the model frequently misclassified data, particularly for the minority class. SVM and KNN also exhibited suboptimal performance, with accuracy rates of 58.44% and 63.64%, respectively. Furthermore, a significant performance gap between training and test data was observed, especially in Decision Tree and Random Forest, which achieved 100% accuracy on training data. This suggests the presence of overfitting, where the models adapt too closely to the training data, reducing their ability to generalize effectively to test data.

Table 4. Model Performance on Imbalanced Data

Method Machine Learning	Test				Train			
	Acc (%)	Pre (%)	Rec (%)	F1-Score (%)	Acc (%)	Pre (%)	Rec (%)	F1-Score (%)
SVM	58.44	35.37	58.44	44.07	63.52	65.50	63.52	51.87
Naïve Bayes	54.55	77.41	54.55	52.61	98.70	98.83	98.70	98.72
KNN	63.64	58.64	63.64	54.27	75.57	80.69	75.57	71.91
Decision Tree	79.22	80.57	79.22	78.91	100	100	100	100
Random Forest	75.32	80.33	75.32	72.41	100	100	100	100

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2) Model Performance on Balanced Data

After balancing the data, there was a significant improvement in model performance. The accuracy of SVM increased from 58.44% to 71.43%, while the F1-score improved from 44.07% to 67.91%. This indicates that SVM performs more effectively on balanced data, whereas class distribution is more even. KNN also improved, with accuracy rising from 63.64% to 66.23%. Tree-based models, such as Decision Tree and Random Forest, continued to perform exceptionally well, achieving accuracy above 75%, with Random Forest reaching 79.99%. Meanwhile, Naïve Bayes maintained a relatively low accuracy (54.55%), like its performance on imbalanced data. This suggests that Naïve Bayes is less effective in capturing sentiment patterns than other models.

Table 5. Model Performance on Balanced Data

Method Machine Learning	**Test**				**Train**			
	Acc (%)	**Pre (%)**	**Rec (%)**	**F1-Score (%)**	**Acc (%)**	**Pre (%)**	**Rec (%)**	**F1-Score (%)**
SVM	71.43	75.10	71.43	67.91	77.42	86.54	77.42	77.67
Naïve Bayes	54.55	77.41	54.55	52.61	99.28	99.30	99.28	99.28
KNN	66.23	68.13	66.23	62.13	92.55	93.60	93.55	93.55
Decision Tree	75.32	77.02	75.32	75.20	100	100	100	100
Random Forest	79.99	82.99	79.22	77.14	100	100	100	100

3) Comparison of Model Performance Between Imbalanced and Balanced Data
   a) Decision tree-based models (Decision Tree & Random Forest) performed best in both scenarios.
   b) SVM and KNN showed a significant improvement after data balancing, indicating that these models are more sensitive to data imbalance.
   c) Naïve Bayes consistently exhibited lower performance than other methods on imbalanced and balanced data.
   d) Overfitting was more evident in Decision Tree and Random Forest models, particularly due to accuracy reaching 100% on training data but lower on test data.

Overfitting occurs when a Machine Learning model adapts too closely to patterns in the training data, leading to a loss of generalization ability on test data [33]. In this study, overfitting is observed in Decision Tree and Random Forest models, where the training data accuracy reaches 100% but is significantly lower on test data. This is due to the high model complexity, which causes it to memorize patterns in the training data, including noise or specific details that do not always appear in new data. Additionally, many features after the text extraction process using CountVectorizer can exacerbate overfitting, especially if regularization

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

techniques or pruning are not applied to tree-based models. As a result, although the model appears highly accurate on training data, its performance declines on test data due to its failure to recognize general patterns in previously unseen data.

3.5 Discussion

This study successfully conducted sentiment analysis on public opinions regarding the 2024 South Sumatra Governor Election, based on data collected from Twitter and YouTube. Through data crawling on Twitter, 140 data points were obtained, while scraping from YouTube resulted in 384 data points after cleaning. The sentiments in these data were classified into three categories: positive, neutral, and negative, using the sentiment_analysis_lexicon_indonesia method. Overall, the analysis results indicate that neutral sentiment dominates both Twitter data (50%) and YouTube data (55%), followed by positive sentiment, while negative sentiment has the smallest proportion. This suggests that the public tends to express informative or descriptive opinions rather than strong support or sharp criticism toward the gubernatorial candidates.

In terms of model performance, this study compares five Machine Learning methods: SVM, Naïve Bayes, KNN, Decision Tree, and Random Forest, on imbalanced and balanced data. The evaluation results show that Decision Tree and Random Forest achieve the highest accuracy on imbalanced data, at 79.22% and 75.32%, respectively. However, both models exhibit overfitting, as indicated by their training data accuracy reaching 100%. After data balancing, the performance of SVM and KNN improves, with accuracy increasing to 71.43% and 66.23%, respectively, suggesting that these models are more sensitive to balanced data distribution. Conversely, Naïve Bayes maintains a low accuracy of 54.55%, indicating that this method is less effective in handling sentiment analysis in the context of the gubernatorial election.

Overall, the results of this study indicate that public opinion regarding the 2024 South Sumatra Gubernatorial Election tends to be more neutral than positive or negative polarity. Additionally, this study proves that selecting the appropriate Machine Learning method and data balancing techniques significantly impacts the accuracy of sentiment analysis. Decision tree-based models (Decision Tree and Random Forest) demonstrate high performance but are prone to overfitting, whereas SVM and KNN exhibit greater stability after data balancing. With these findings, this research can serve as a reference for developing more accurate public opinion monitoring systems, particularly in Indonesia's political and election sentiment analysis. However, this study also has some limitations that can be improved in future research. One of the main limitations is using the lexicon-based method (sentiment_analysis_lexicon_indonesia) for data labeling, which is less accurate in handling sarcasm, idioms, and more complex sentence contexts.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Additionally, this study only employs CountVectorizer as the feature extraction method, which relies solely on word frequency without considering the semantic relationships between words in a sentence. Some models, particularly Decision Tree and Random Forest, also exhibit signs of overfitting, necessitating further optimization strategies.

Table 6. Comparison Model Performance

Author/year	Sumber	Data	Method	Acc
Putu Pebri Armaeni et al. (2024) [34]	YouTube	7000	Naïve Bayes Decision Tree	73.96% 74.71%
Alisya Mutia Mantika et al. (2024) [35]	Twitter	1573	Naive Bayes	59%
Our Study	YouTube Twitter	538	SVM Naïve Bayes KNN Decision Tree Random Forest	71.43 54.55 66.23 75.32 79.99

Table 6 provides a comparison of the performance of various machine learning models on balanced data from previous studies [XX], [YY] as well as our own study. The table details the data source (YouTube or Twitter), the number of data points used, the methods applied, and the corresponding accuracy achieved by each method. This comparison highlights how different models perform under balanced conditions and underscores the potential impact of data balancing techniques on improving model accuracy.

Future research can be further developed by utilizing deep learning methods such as Word2Vec, FastText, or BERT, which can capture word meanings in a broader context. Additionally, the dataset's quality can be improved by involving human annotators in the labeling process to ensure sentiment label accuracy. From a modeling perspective, regularization techniques, hyperparameter tuning, and more complex ensemble learning methods can be applied to address overfitting and enhance model generalization to new data. With these advancements, future studies are expected to produce more accurate sentiment analysis models that can be widely used for monitoring public opinion, particularly in politics and elections.

4. CONCLUSION

This study comprehensively analyzes public sentiment toward the 2024 South Sumatra Gubernatorial Election using five machine learning methods—SVM, Naïve Bayes, KNN, Decision Tree, and Random Forest—evaluated on both imbalanced and balanced datasets. Data were collected from Twitter (140 data

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

points) and YouTube (384 data points) via crawling and scraping, followed by preprocessing and feature extraction using CountVectorizer with N-Gram. The results indicate that neutral sentiment dominates the discourse on both platforms, with Twitter showing 50% and YouTube 55% neutral sentiments. In contrast, positive sentiment is more prevalent than negative sentiment, suggesting that the public tends to express informative opinions rather than overt support or criticism. Regarding model performance, Decision Tree and Random Forest achieved the highest accuracies on imbalanced data (79.22% and 75.32%, respectively) but were prone to overfitting, as evidenced by their near-perfect training performance. Conversely, after applying data balancing techniques, SVM and KNN demonstrated more stable improvements, with accuracies increasing to 71.43% and 66.23%. The key strengths of this study include integrating multiple data sources and machine learning approaches, along with a thorough evaluation of imbalanced and balanced scenarios. However, limitations such as the relatively small sample size and the overfitting tendencies of tree-based models warrant caution. Future research should focus on incorporating larger datasets and exploring advanced techniques, such as deep learning, to further enhance the robustness and applicability of sentiment analysis in political contexts.

REFERENCES

[1] H. M. Duryat and M. Pd, *Indramayu: Menuju Kontestasi Pilkada 2024, Problem Kepemimpinan, Demokratisasi dan Pembangunan Berkelanjutan*. Penerbit Adab, 2024.

[2] S. N. Rahim, H. N. Shabrina, R. Salsabila, S. Hanum, N. A.-R. Zemlya, and A. Rahman, "Peran buzzer di media sosial dalam membentuk opini kebijakan publik di masyarakat pada Pemilu 2024," *PubBis: J. Pemikir. Penelit. Adm. Publik Adm. Bisnis*, vol. 8, no. 2, pp. 147–158, 2024.

[3] A. Karim, *Big Data Analytics: Analisis Sentimen Netizen di Era Media Baru*. Penerbit NEM, 2025.

[4] A. J. Sutan, "Media sosial di masa pandemi: Media sosial digunakan untuk menolak diskriminasi rasial kasus kampanye #AsianLivesMatter di Amerika Serikat," *Analisis*, p. 93.

[5] R. Safitri, N. Alfira, D. Tamitiadini, W. W. A. Dewi, and N. Febriani, *Analisis Sentimen: Metode Alternatif Penelitian Big Data*. Universitas Brawijaya Press, 2021.

[6] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022.

[7] A. Peivandizadeh *et al.*, "Stock market prediction with transductive long short-term memory and social media sentiment analysis," *IEEE Access*, 2024.

p-ISSN: 2656-5935 http://journal-isi.org/index.php/isi e-ISSN: 2656-4882

[8] M. R. Pavan Kumar and P. Jayagopal, "Context-sensitive lexicon for imbalanced text sentiment classification using bidirectional LSTM," *J. Intell. Manuf.*, vol. 34, no. 5, pp. 2123–2132, 2023.
[9] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, 2024.
[10] A. R. Hakim, W. Gata, A. Z. P. Widodo, O. Kurniawan, and A. R. Syarif, "Analisis perbandingan algoritma machine learning terhadap sentimen analis pemindahan ibu kota negara," *J. JTIK (J. Teknol. Inf. Komun.)*, vol. 7, no. 2, pp. 179–185, 2023.
[11] R. Aryanti, T. Misriati, and A. Sagiyanto, "Analisis sentimen aplikasi Primaku menggunakan algoritma Random Forest dan SMOTE untuk mengatasi ketidakseimbangan data," *J. Comput. Syst. Informatics (JoSYC)*, vol. 5, no. 1, pp. 218–227, 2023.
[12] M. M. Effendy, T. E. Sutanto, and M. Liebenlito, "Efektivitas variabel demografi pengguna Twitter dalam prediksi Pilpres Indonesia 2014 dan 2019," *Indones. J. Comput. Sci.*, vol. 12, no. 6, 2023.
[13] L. A. Andika, P. A. N. Azizah, and R. Respatiwulan, "Analisis sentimen masyarakat terhadap hasil quick count pemilihan presiden Indonesia 2019 pada media sosial Twitter menggunakan metode Naïve Bayes classifier," *Indones. J. Appl. Stat.*, vol. 2, no. 1, pp. 34–41, 2019.
[14] A. Halim *et al.*, "Klasifikasi sentimen masyarakat di Twitter terhadap Prabowo Subianto sebagai bakal calon presiden 2024 menggunakan M-KNN," *J. Inf. Syst. Res. (JOSH)*, vol. 5, no. 1, pp. 202–212, 2023.
[15] M. K. Anam, B. N. Pikir, M. B. Firdaus, S. Erlinda, and A. Agustin, "Penerapan Naïve Bayes classifier, K-Nearest Neighbor (KNN) dan decision tree untuk menganalisis sentimen pada interaksi netizen dan pemerintah," *MATRIK: J. Manaj. Tek. Inform. Rekayasa Komput.*, vol. 21, no. 1, pp. 139–150, 2021.
[16] D. Alita and A. R. Isnain, "Pendeteksian sarkasme pada proses analisis sentimen menggunakan Random Forest classifier," *J. Komputasi*, vol. 8, no. 2, pp. 50–58, 2020.
[17] W. W. H. Cholil, F. Panjaitan, F. Ferdiansyah, A. Arista, R. Astriratma, and T. Rahayu, "Comparison of machine learning methods in sentiment analysis PeduliLindungi applications," in *Proc. 2022 Int. Conf. Informatics, Multimedia, Cyber Inf. Syst. (ICIMCIS)*, Nov. 2022, pp. 276–280, IEEE.
[18] E. Setiani and W. Ce, "Text classification services using Naïve Bayes for Bahasa Indonesia," in *Proc. 2018 Int. Conf. Inf. Manag. Technol. (ICIMTech)*, IEEE, 2018, pp. 361–366.
[19] G. Kanugrahan and A. F. Wicaksono, "Sentiment analysis of face-to-face learning during COVID-19 pandemic using Twitter data," in *Proc. 2021 8th Int. Conf. Adv. Informatics: Concepts, Theory Appl. (ICAICTA)*, IEEE, 2021, pp. 1–6.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

[20] I. K. A. B. Artana, G. A. Pradnyana, and I. G. M. Darmawiguna, "Analisis sentimen Twitter untuk menilai kesiapan pembelajaran tatap muka terbatas dengan Inset Lexicon dan Levenshtein Distance," *J. Pendidik. Teknol. Kejuruan*, vol. 20, no. 2, pp. 200–209, 2023.

[21] M. K. Anam, "Penerapan metode support vector machine untuk analisis sentimen terhadap produk skincare," *Indones. J. Comput. Sci.*, vol. 13, no. 1, 2024.

[22] V. Bonta, N. Kumaresh, and N. Janardhan, "A comprehensive study on lexicon-based approaches for sentiment analysis," *Asian J. Comput. Sci. Technol.*, vol. 8, no. S2, pp. 1–6, 2019.

[23] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, "Framework for improved sentiment analysis via random minority oversampling for user tweet review classification," *Electronics (Basel)*, vol. 11, no. 19, p. 3058, 2022.

[24] M. Fachrie, A. Musdholifah, and S. Hartati, "Improving sentiment analysis performance on imbalanced dataset using data resampling and statistical feature selection," in *Proc. 2024 8th Int. Conf. Inf. Technol. (InCIT)*, Nov. 2024, pp. 272–277, IEEE.

[25] A. Miftahusalam, A. F. Nuraini, A. A. Khoirunisa, and H. Pratiwi, "Perbandingan algoritma Random Forest, Naïve Bayes, dan support vector machine pada analisis sentimen Twitter mengenai opini masyarakat terhadap penghapusan tenaga honorer," in *Seminar Nasional Official Statistics*, vol. 2022, no. 1, pp. 563–572, Nov. 2022.

[26] I. N. Husada and H. Toba, "Pengaruh metode penyeimbangan kelas terhadap tingkat akurasi analisis sentimen pada tweets berbahasa Indonesia," *J. Tek. Inform. Sist. Informasi*, vol. 6, no. 2, 2020.

[27] B. B. Baskoro, I. Susanto, and S. Khomsah, "Analisis sentimen pelanggan hotel di Purwokerto menggunakan metode Random Forest dan TF-IDF (studi kasus: ulasan pelanggan pada situs TRIPADVISOR)," *J. Informatics Inf. Syst. Softw. Eng. Appl. (INISTA)*, vol. 3, no. 2, pp. 21–29, 2021.

[28] O. I. Gifari, M. Adha, I. R. Hendrawan, and F. F. S. Durrand, "Analisis sentimen review film menggunakan TF-IDF dan support vector machine," *J. Inf. Technol.*, vol. 2, no. 1, pp. 36–40, 2022.

[29] I. Saputra, R. S. A. Pambudi, H. E. Darono, F. Amsury, M. R. Fahdia, B. Ramadhan, and A. Ardiansyah, "Analisis sentimen pengguna marketplace Bukalapak dan Tokopedia di Twitter menggunakan machine learning," *Faktor Exacta*, vol. 13, no. 4, pp. 200–207, 2021.

[30] G. A. Buntoro, R. Arifin, G. N. Syaifuddiin, A. Selamat, O. Krejcar, and F. Hamido, "The implementation of the machine learning algorithm for the sentiment analysis of Indonesia's 2019 presidential election," *IIUM Eng. J.*, vol. 22, no. 1, pp. 78–92, 2021.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

[31] C. A. N. Agustina, R. Novita, and N. E. Rozanda, "The implementation of TF-IDF and Word2Vec on booster vaccine sentiment analysis using support vector machine algorithm," *Procedia Comput. Sci.*, vol. 234, pp. 156–163, 2024.
[32] I. Lazrig and S. L. Humpherys, "Using machine learning sentiment analysis to evaluate learning impact," *Inf. Syst. Educ. J.*, vol. 20, no. 1, pp. 13–21, 2022.
[33] P. Rahayu, I. G. I. S. Wibawa, S. Suryani, A. Surachman, A. Ridwan, I. G. M. Darmawiguna, M. N. Sutoyo, I. Slamet, S. Harlina, and I. M. D. Maysanjaya, *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia, 2024.
[34] P. P. Armaeni, I. K. A. G. Wiguna, and W. G. S. Parwita, "Sentiment analysis of YouTube comments on the closure of TikTok Shop using Naïve Bayes and decision tree method comparison," *J. Galaksi*, vol. 1, no. 2, pp. 70–80, 2024.
[35] A. M. Mantika, A. Triayudi, and R. T. Aldisa, "Sentiment analysis on Twitter using Naïve Bayes and logistic regression for the 2024 presidential election," *SaNa: J. Blockchain, NFTs Metaverse Technol.*, vol. 2, no. 1, pp. 44–55, 2024.