Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 9, No. 1, March 2026, pp. 72 – 82
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Analyzing Opinion Polarization on Joko Widodo's Diploma Using Machine Learning

[1*]Julianti, [2]Farid Wajidi, [3]Musawwir
[1,2,3]Department of Informatics Engineering, University of West Sulawesi, Majene
Email: [1] juliantijuly39@gmail.com, [2]faridwajidi@unsulbar.ac.id, [3]musawwir@unsulbar.ac.id

Article Info
Article history:
Received Nov 27th, 2025
Revised Feb 24th, 2026
Accepted Mar 05th, 2026

Keyword:
Machine Learning
Naïve Bayes
Public Opinion Polarization
Random Forest
Sentiment Analysis

ABSTRACT
The authenticity of former President Joko Widodo's diploma has become a hot topic on the digital space, especially in the comments section of Kompas TV's YouTube channel. The wide diversity of opinions reflects a polarization of public opinion that is worth further analysis. Given the large volume of text data from public comments, manual analysis is ineffective; a technology-based approach is needed to systematically group opinions. Therefore, this study was conducted to analyze public opinion polarization using a machine learning approach. Two classification algorithms, Naive Bayes and Random Forest, were used to distinguish between pro and con public comments on the issue. Data were obtained through an automated collection process (web scraping), followed by text pre-processing and TF-IDF (Term Frequency–Inverse Document Frequency) word weighting. The test results showed that the Random Forest algorithm performed best, with an accuracy of 91%, while Naïve Bayes achieved only 74%. This shows that the Random Forest method is more effective than the Naïve Bayes approach in detecting unstructured text patterns. This study concludes that machine learning can be used effectively to identify trends in public opinion on social media and can serve as a basis for further research using word embedding and deep learning models.

Corresponding Author:
Julianti,
Department of Informatics Engineering,
University of West Sulawesi,
Jalan Prof. Dr. Baharuddin Lopa, SH, Talumung, Majene, West Sulawesi
Email: juliantijuly39@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v9i1.38616

1. INTRODUCTION

Public opinion plays an increasingly important role in shaping socio-political dynamics amid the development of digital technology and easy access to information. The rise of social media platforms, which serve not only as a means of entertainment but also as an open arena for the public to express their opinions, criticisms, and even support for various national issues. YouTube is a video-based social media platform that is used not only for entertainment, but also as a digital public space where information and debates on social and political issues are shared. Content such as news, interviews, and public debates, including discussions about public figures, are widely shared and often elicit broad public response. The availability of the comment feature on YouTube can help identify trends in public opinion, whether an issue receives more support or criticism [1] .

Indonesia is one of the countries with the most YouTube users in the world. According to rri.co.id, based on a report by We Are Social, Indonesia is one of the countries with the most YouTube users at the beginning of 2025. Indonesia recorded 143 million YouTube users at the beginning of this year, contributing 5.65% of the total global users. One of the issues currently being widely debated is the authenticity of former Indonesian President Joko Widodo's diploma. Various YouTube videos have attracted many viewers and generated numerous comments reflecting their opinions and feelings about the controversy.

Allegations regarding the authenticity of former President Joko Widodo's diploma have become a controversial topic attracting public attention, especially after his term ended. Although Gadjah Mada University and government agencies have issued official clarifications, some people remain skeptical of the document's authenticity. The debate then spread to social media, especially YouTube, where comments on the videos ranged from support to rejection of the allegations, reflecting the polarization of public opinion in the digital space. Given the large amount of comment data and the unstructured form of the text, the process of identifying public opinion manually is ineffective. Therefore, an information technology-based approach, specifically sentiment analysis aided by Machine Learning Algorithms, is needed to systematically and measurably classify public opinion into pro and con categories [2],[3].

Several previous studies were used as supporting material in this study. In a study analyzing sentiment related to vote-counting results that caused discrepancies in the Sirekap 2024 application, conducted by [4]. The researchers used Support Vector Machine, Naïve Bayes, and Random Forest classification methods to classify the data. This study began by scraping data using a Python program on the Google Colab application with the Google Play Scraper technique on the Android Play Store application Sirekap 2024. Next, data preprocessing was performed to clean and simplify the text, making it easier for the sentiment labeling algorithm to process using the TextBlob technique. Support Vector Machine classification yielded 82%, Naive Bayes yielded 71%, and Random Forest yielded 81%. The best performance among the three classification models for sentiment identification was achieved by the Support Vector Machine, with an accuracy of 82%, precision of 82%, recall of 82%, and an F1-score of 82%.

Another study conducted by [5] analyzed comments from the 2019 presidential election debate YouTube channel. This study balanced the amount of positive and negative data using oversampling and extracted features with the Skip-gram type of Word2Vec, where each comment feature weight was calculated using an average-based approach. For the classification model, the study used Random Forest with different epoch and window parameters. Model testing yielded accuracies of 88.77% to 89.05%.

Another study analyzed YouTube comments on the topic of Islamophobia, conducted by [6]. The study used 1,000 Indonesian-language YouTube comments related to videos depicting bombings or terrorist attacks. The study noted that the Random Forest algorithm performed sentiment classification quickly, resulting in an accuracy rate of 79%. The data used was collected using scrapestom and labeled manually using the crowdsourcing method. Next, the comment data will be preprocessed and word-weighted using TF-IDF, then classified using the Random Forest algorithm.

The polarization of public opinion in the digital space is an increasingly complex phenomenon as social media becomes more prevalent as a space for political discourse. The issue of the authenticity of former President Joko Widodo's diploma has not only become a personal debate, but has also developed into a public discourse that has triggered fragmentation of public opinion. In the context of digital political communication, this polarization can affect institutional legitimacy and public perceptions of political figures. Therefore, data-based analysis of public opinion dynamics is important for objectively and measurably understanding patterns of public attitudes.

Previous studies have examined sentiment analysis on political issues, such as election debates, government policies, and perceptions of other public figures. However, most of these studies focus on general sentiment classification (positive, negative, neutral) without specifically examining the polarization of opinion based on a public figure's academic legitimacy. In addition, there has not been much research that comparatively evaluates the performance of classification algorithms in the context of imbalanced class distribution on controversial issues such as this. The latest research [7], analyzes sentiment toward the 2024 Indonesian presidential candidates using SVM and Random Forest, with a focus on positive and negative classification. Meanwhile, the research [8] analyzes public sentiment toward President Joko Widodo's performance using the LSTM-based Recurrent Neural Network (RNN) method. These studies demonstrate the effectiveness of the deep learning approach in understanding public opinion on social media. However, the research is still limited to general sentiment classification and has not examined opinion polarization around specific issues or the problem of class imbalance in the context of controversial issues.

Based on this gap, this study aims to identify patterns of polarization in public opinion on the issue of diploma authenticity through pro and con classification, as well as to compare the performance of the Naïve Bayes and Random Forest algorithms in classifying public opinion based on YouTube comment data. Thus, this study is expected to contribute methodologically to machine learning-based political sentiment analysis and enrich studies on the dynamics of public opinion in the Indonesian digital space.

2. RESEARCH METHOD

This study will discuss comments or opinions regarding the debate on the authenticity of former Indonesian President Joko Widodo's diploma on the Kompas TV YouTube channel. The classification methods used are Naive Bayes and Random Forests for sentiment analysis. This research is expected to be

useful for classifying public comments or opinions between those for and against the authenticity of the diploma. In addition, this study also aims to determine which algorithm has the best accuracy in performing classification. Figure 1 shows the workflow of the attached research stages.

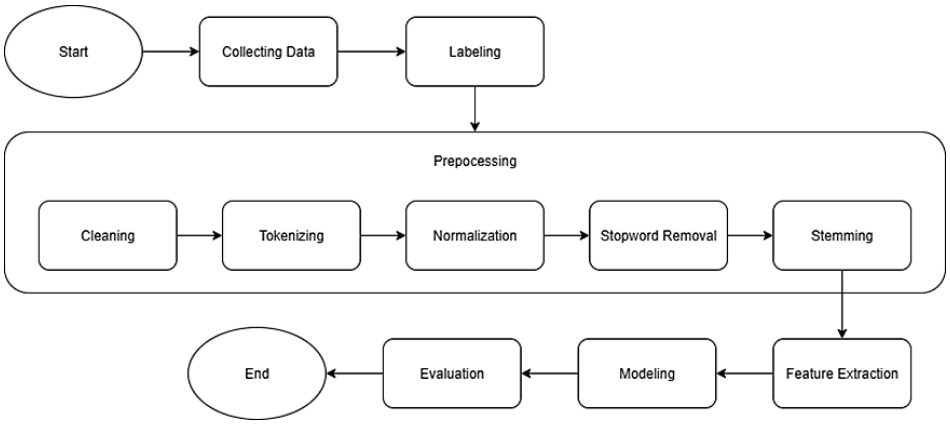


Figure 1. Research Method Flowchart

2.1. Collecting Data

The first step was to collect data obtained from social media on the Kompas TV YouTube channel using the keyword "Roy Suryo's debate on Jokowi's diploma." Data collection was carried out using scraping techniques on Google Colab with Python programming, assisted by the YouTube API, which is available for free. The scraping resulted in 6,726 comments, some examples of which are shown in Table 1.

Table 1. The Example of Comments

Index	Comments
1	Kalaw ijazah terbukti asli, penjarakan jgn kasih ampun orng2 itu khususnya roy, supaya gk ad lg orng2 yg nyer2 kdepannya
2	Pak Joko sedang tdk jujur UGM sedang mempertaruhkan nama baiknya.
3	Pencemaran nama baiklah, kenapa karena ijasah asli koq di gembor²kan palsu..Mau enggak ijasah mas suryo di bilang palsu
4	Pekerjaan yg sangat capek, yaa bohong teroooosss 😂
5	Jujur itu simpel..kebohongan sll berbelit2 penuh drama sampai bikin gaduh.salam akal sehat

2.2. Labelling

Before labeling, comments are filtered so that only relevant comments are processed in the next stage of analysis. The comment filter resulted in only 2161 comments. Labeling was done using Lexicon labeling, which produced pro and con comments. Sentiment classification using a lexical approach with a valence-oriented dictionary and sentiment analyzer (VADER) is an analysis method that uses a dictionary of positive, negative, or neutral words developed for informal texts such as social media. VADER enables computers to understand sentiment based on the polarity values of words in a lexicon, with emotional intensity adjustable through word reinforcement or punctuation. This approach facilitates the analysis of customer opinions about products or services, as it can process informal language, emoticons, and punctuation, making it an effective solution for accurate social media data analysis[9]. Although lexicon approaches like VADER conceptually produce general polarity categories such as positive, negative, and neutral, in this study, the polarity results were not used directly as final labels. Instead, the polarity results were converted into pro and con categories based on the issue being analyzed, namely the debate over the authenticity of former President Joko Widodo's diploma.

2.3. Preprocessing

In the data preprocessing stage, raw data will be converted to a more structured, cleaner format to improve analysis efficiency and enhance model performance. Data preprocessing is very important for improving data quality before training the model, as it removes irrelevant information, addresses data deficiencies, and even prepares relevant features for the next stage. Here are some stages in data preprocessing [10].

1. Cleaning

   The first step in pre-processing data is to clean the text data of irrelevant elements such as mentions, hashtags, emoticons, links, numbers, and punctuation marks. At this stage, all text will also be converted to lowercase letters [11].

2. Tokenizing
   This process divides the review text into smaller units called tokens, usually in the form of words or short phrases. After the text is cleaned, at this stage it will be divided into smaller units in the form of tokens or short phrases [11].

3. Normalization
   At this stage, non-standard words will be changed to standard words, such as "aja" to "saja" and "ngapain" to "kenapa" [4].

4. Stopword Removal
   At this stage, words that do not affect the classification process will be removed, such as "dan," "ke," "atau," "dari," "siapa," and so on [12].

5. Stemming
   The final stage of data preprocessing is to select words that have conjunctions, affixes, pronouns, and verbs as root words, by removing their prefixes or suffixes [12].

2.4. Feature Extraction

After going through the data pre-processing stage, feature extraction is carried out using word weighting, namely TF-IDF, where calculations give weight to each word that is relevant to how often the word appears in a particular document[11] . This method combines two important concepts: Term Frequency (TF), which measures how often a word appears, with a higher value for more frequent words, and Inverse Document Frequency (IDF), which measures how important a word is in the entire corpus, with less frequent words considered more important[13] ,[14] . The formula for calculating TF-IDF is presented in Equation 1.

$$TFIDF_{std}(t)^{t} = tf_{d} \times \log log \frac{N}{df}_{d}^{t} \quad (1)$$

Description:

tf^t : frequency of occurrence of term *t* in document *d*
N^d : total number of documents
df^t : the number of documents containing term *t*

2.5. Modeling

The processed data, extracted features, will then be trained with a model using the Naïve Bayes and Random Forest methods.

1. Naïve Bayes
   The Naïve Bayes algorithm is based on Bayes' theorem. In sentiment analysis, the Naïve Bayes classification technique is among the most popular methods. This classification technique has a strong basis in Bayes' theorem and performs comparably to decision tree and neural network methods. In its implementation, this algorithm considers the weights of conditional probabilities, with an objective function modeled after the Naïve Bayes classifier structure. Optimal weight determination is typically performed using local optimization methods, such as the quasi-Newton method. In several studies, Naïve Bayes has been used as a starting point for the development of new methods[15], [14], [16]. The mathematical equation for Bayes' theorem is presented in equation 2.

$$P(X) = \frac{P(X|H)}{P(X)} P(H) \quad (2)$$

Description:

X : feature data (TF-IDF results of comments).
H : class (positive/negative/neutral).
$P(H)$: prior probability of a class.
$P(X|H)$: probability of a feature appearing in a particular class.
$P(X)$: the probability of a comment belonging to a specific class (prediction).

2. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to produce a more accurate classification model. It works by building multiple decision trees from random samples of the dataset, then generating predictions based on majority voting for classification or averaging for regression. Thus, Random Forest is able to reduce the risk of overfitting that often occurs in single decision trees [17], [13], [18]. The advantages of the Random Forest method can be seen from its ability to produce high accuracy, resistance to outliers and noise, and faster performance compared to bagging and boosting methods [19]. The mathematical equations for Random Forest are presented in equations 3 and 4 [20].

$$Entropy(Y) = -\sum_{c \epsilon C} p(c|Y)\, log_2\, p(c|Y) \tag{3}$$

$$InformationGain(Y, a) = Entropy(Y) - \sum_{values(a)} \frac{Y_p}{Y_v}\ Entropy(Y_p) \tag{4}$$

Description:

Y	: Data collection.
c	: Possible classes.
$p(c\|Y)$	: proportion of instances in Y that belong to class c.
a	: The attribute being tested.
Values(a)	: all possible values of attribute a.
Yv	: a subset of Y that has attribute value a = v.
$\|Y_v\|/\|Y\|$	: the proportion of that subset relative to the whole.
$Entropy(Y_p)$	: entropy of that subset.

2.6. Evaluation

At this stage, the trained model is then used to make predictions on the test data. The prediction results are then evaluated using a confusion matrix, which calculates four main categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In addition to the confusion matrix, it serves as a tool for calculating important metrics such as accuracy, precision, recall, and F1-Score. Accuracy is defined as the ratio of correct predictions to the total number of predictions for all data. The accuracy equation is presented in equation 5. Precision is the value obtained from the comparison between true positives and the number of predictions of data that are assessed as positive, which is presented in equation 6. Recall is the value obtained from the comparison between true positives and the number of data that are true positives, also known as true positive rate (TPR) or sensitivity, presented in equation 7. F1-Score is the comparison of the average value between precision and recall, presented in equation 8 [21],[22] .

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

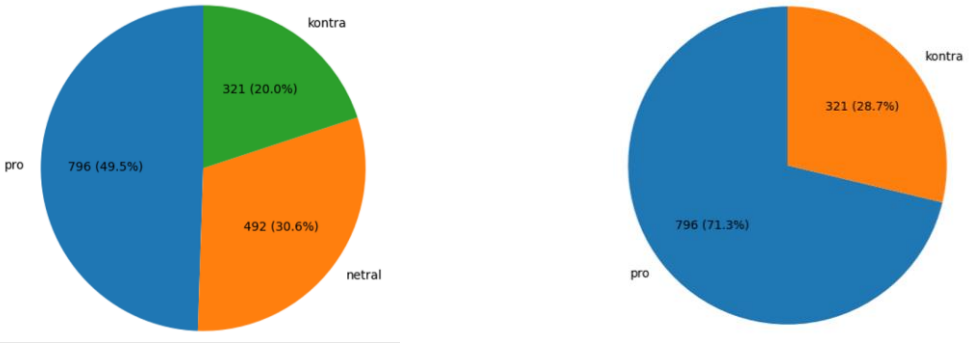
$$F - 1\ Score = 2 * \frac{Recall * Precission}{Recall + Precission} \tag{8}$$

3. RESULTS AND ANALYSIS

The raw data obtained from crawling consists of 5,017 user comments on the issue of the authenticity of Joko Widodo's diploma. The raw data will first be processed to be grouped into three classes: pro, con, and neutral through lexicon labeling. After class division, the neutral class will be removed as shown in Figure 2. Some examples of the labeling process results are shown in Table 2.

Table 2. The Example of Labeling

Index	Comments	Label
1	Ijazah Palsu dibela-belain direkayasa disembunyikan sampai ke ujung dunia tetap Ijazah Palsu	Kontra
2	Fix palsu,kalau asli tinggal tunjukan aja ke media ngapain takut haahaahaa 😂😂	Kontra
3	Kok bulet ya. Tinggal tunjungin aj ijazah asli nya kn beresssssss engak ribetttttttt bikin rameee gaduh	Pro
4	Tunjukin aja pa ijazah asli bpk selesai pa...kita pakai paham pancasila aja biar adem rakyat mu...jangan pakai paham.yg lain ...	Pro
5	Ijazahnya itu jelas² palsu boooosssss	Kontra



(a) (b)

Figure 2. Data in the label class

(pros) (cons)

Figure 3. Word cloud in label class

Figure 2 shows the results of removing the neutral class from the clustered data. It can be seen that the pro class accounts for 71.3% of the data and the contra class accounts for 28.7%. This data will be processed in the data pre-processing stage, the results of which are presented in Table 3. Figure 3 displays a *word cloud* visualization based on the distribution of words in each label class, namely the pro and contra classes, where the size of the words represents their frequency of occurrence in the data. The difference in word occurrence patterns in the two classes shows different lexical characteristics and supports the results of the sentiment labeling that was done earlier. The raw data was previously processed by converting everything to lowercase letters, removing all irrelevant elements, dividing it into smaller units in the form of tokens, normalizing all non-standard words, removing words that had no effect on the classification process, and removing all affixes to leave only the root words.

Table 3. The Example of Text Preprocessing

Stages	Result
Cleaning	fix palsu kalau asli tinggal tunjukan aja ke media ngapain takut haahaahaa
Tokenizing	‘fix’, ‘palsu’, ‘kalau’, ‘asli’, ‘tinggal’, ‘tunjukan’, ‘aja’, ‘ke’, ‘media’, ‘ngapain’, ‘takut’, ‘haahaahaa’
Normalization	‘fix’, ‘palsu’, ‘kalau’, ‘asli’, ‘tinggal’, ‘tunjukkan’, ‘saja’, ‘ke’, ‘media’, ‘kenapa’, ‘takut’, ‘hahaha’
Stopword Removal	‘fix’, ‘palsu’, ‘kalau’, ‘asli’, ‘tinggal’, ‘tunjukkan’, ‘media’, ‘kenapa’, ‘takut’
Stemming	‘fix’, ‘palsu’, ‘kalau’, ‘asli’, ‘tinggal’, ‘tunjuk’, ‘media’, ‘kenapa’, ‘takut’

Next, these texts will undergo a feature extraction process involving word weighting or themes represented in vector form to facilitate the classification process. Several examples can be seen in Table 4.

Before the classification process was carried out, the initial distribution of lexicon-based labeling data showed class imbalance, with 796 data points in the pro class and 321 data points in the contra class. This imbalance can lead to model bias towards the majority class. Therefore, the undersampling technique was applied to the pro class so that the number of data in both classes became balanced, with 321 data points each.

Table 4. Examples of TF-IDF word weighting results

Word	TF-IDF
Ijazah	0,93
Tidak	0,96
Palsu	1,03
Orang	0,00
Benar	1,07

3.1. Naive Bayes

Testing using the Naive Bayes algorithm showed an accuracy rate of 74.42%. The precision, recall, and f1-score values for both classes were in the range of 0.73 to 0.76, indicating a relatively balanced performance distribution. This indicates that the model is able to classify pro and con classes ( ) quite consistently. However, the accuracy rate obtained is still in the middle category, so the generalization ability of the model can be said to be not yet optimal. Accuracy evaluation and confusion matrix of the naive bayes method and classification report of naïve bayes can be seen in Figure 4 and Table 5

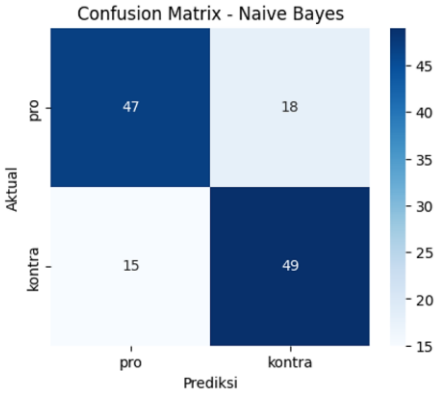


Figure 4. Accuracy evaluation and confusion matrix of the Naive Bayes method

Table 5. Classification report of Naïve Bayes

	Precision	Recall	F1-Score	Support
Cons	0.76	0.72	0.74	65
Pros	0.73	0.77	0.75	64
Accuracy			0.74	129
Macro average	0.74	0.74	0.74	129
Weighted average	0.74	0.74	0.74	129

3.2. Random Forest

Unlike Naive Bayes, the Random Forest algorithm shows superior performance. This model achieves an accuracy of 90.70%, with precision and recall values for both classes ranging from 0.86 to 0.95. The f1-score obtained ranges from 0.90 to 0.91. In addition, the similarity of the macro average and weighted average values indicates that the model has good performance stability in each class after the data balancing process. Overall, these results show that Random Forest is more effective in classification than Naive Bayes on datasets that have undergone undersampling. Accuracy evaluation and confusion matrix of the random forest method and classification report of random forest can be seen in Figure 5 and Table 6.

3.3. Model Performance Comparison

The visual comparison of models in this study aims to illustrate the accuracy levels achieved by the two models used. Figure 6 and Table 7 show a comparison of the performance of the Naive Bayes and Random Forest models based on accuracy values in the dataset that has undergone class balancing. The evaluation results show that the Naive Bayes model achieved an accuracy of 0.74 (74%), while Random Forest achieved an accuracy of 0.91 (91%).

The visualization in Figure 6 shows a significant difference in performance between the two algorithms, with Random Forest demonstrating a higher level of accuracy than Naive Bayes. This finding indicates that Random Forest has a better ability to identify patterns and relationships between features in the data, enabling it to produce more accurate predictions.

On the other hand, although Naive Bayes shows relatively good and stable performance, the probabilistic approach used is simpler than the ensemble mechanism applied in Random Forest. Overall, based on the comparison of accuracy values, Random Forest can be declared as the more optimal model for the classification task in this research dataset.

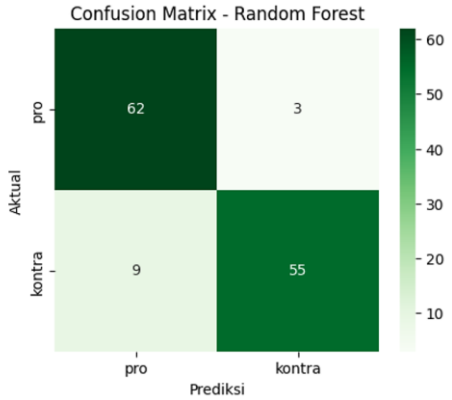


Figure 5. Accuracy evaluation and confusion matrix of the Random Forest method

Table 6. Classification report of Random Forest

	Precision	Recall	F1-Score	Support
Cons	0.87	0.95	0.91	65
Pros	0.95	0.86	0.90	64
Accuracy			0.91	129
Macro average	0.91	0.91	0.91	129
Weighted average	0.91	0.91	0.91	129

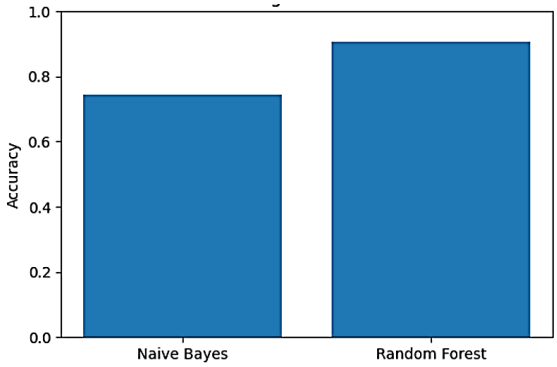


Figure 6. Comparison of the performance of the two models

Table 7. Comparison of model performance

	Model	Accuracy
0	Naive Bayes	0.74
1	Random Forest	0.91

3.4. Discussion

The results of this study indicate that the Random Forest algorithm outperforms Naïve Bayes in classifying pro and con opinions on the issue of diploma authenticity. This difference in performance is in line with the basic characteristics of each algorithm. Naïve Bayes relies on the assumption of feature independence, which limits its ability to represent complex relationships among words in public opinion texts. In contrast, Random Forest, as an ensemble approach, builds a number of decision trees from different data subsets and integrates their results via majority voting. This approach allows the model to capture non-linear patterns and feature interactions more comprehensively.

These findings are in line with research [4] comparing SVM, Naïve Bayes, and Random Forest in sentiment analysis of the Sirekap 2024 application. The study reported that Random Forest (81%) outperformed Naïve Bayes (71%), although SVM achieved the best overall results. A similar trend was also reported by [5], which applied Random Forest with Word2Vec representation and achieved an accuracy of up to 89.05% in the analysis of comments on the 2019 presidential election debate. Additionally, [6] showed that Random Forest was able to achieve 79% accuracy in classifying the sentiment of YouTube comments related to Islamophobia. The consistency of results across these studies reinforces the robustness of Random Forests in processing unstructured social media text rich in linguistic variation.

When compared to study [7], which analyzed the sentiment of the 2024 presidential candidates using SVM and Random Forest, the accuracy achieved by Random Forest in this study (91%) is competitive and even relatively higher than several previous studies. This is likely influenced by the limitation of classifying into two categories (pro and con) after the data balancing process, which narrows the problem's complexity. On the other hand, deep learning approaches such as those proposed in study [8] through the RNN-LSTM model have proven effective in capturing the sequential context of text, but these approaches require greater computational resources and a larger amount of data to achieve optimal performance.

In addition to algorithm selection, data balancing strategies through undersampling also play a role in improving model performance stability. Before balancing, disproportionate class distributions have the potential to cause bias towards the majority class. After the undersampling process is applied, the macro average and weighted average values in Random Forest show consistent results, indicating stable performance in both classes. This finding is in line with a comparative study of machine learning algorithms in sentiment analysis [20], which states that ensemble methods tend to be more robust to data distribution variations than conventional probabilistic methods.

Conceptually, the results of this study reinforce the literature that states that in social media-based political sentiment analysis, ensemble models such as Random Forest are more adaptive to the complexity of informal language, ambiguity of meaning, and the dynamics of public opinion polarization. Thus, this study not only affirms the findings of previous studies but also provides empirical contributions in the context of the issue of academic legitimacy of public figures, which has been limited in comparative studies, particularly when considering the problem of class distribution imbalance.

4. CONCLUSION

This study aims to analyze the polarization of public opinion on the issue of the authenticity of former President Joko Widodo's diploma on the Kompas TV YouTube channel by utilizing a machine learning approach, specifically the Naive Bayes and Random Forest algorithms. Based on the results of the experiment, the Random Forest algorithm showed superior performance with an accuracy rate of 91%, compared to Naive Bayes, which only achieved an accuracy of 74%. These findings are in line with the expectations described in the introduction, where ensemble-based models are expected to produce better generalization in handling complex and unstructured text data.

Empirically, the results of this study reinforce the evidence that the ensemble learning approach is effective in improving the accuracy of sentiment classification in social media data, which is inherently dynamic and full of linguistic context ambiguity. In addition, this study provides a methodological contribution to the application of TF-IDF-based text mining in public opinion analysis, which is relevant to digital data-based socio-political studies.

For further research, it is recommended to develop models through improved feature engineering, the application of word embedding techniques such as Word2Vec or BERT to enrich semantic representation, and the exploration of deep learning models such as LSTM or Transformer. With this direction, it is hoped that sentiment analysis models can achieve higher accuracy while providing a deeper understanding of the dynamics of public opinion in the Indonesian digital space.

ACKNOWLEDGEMENTS

The author would like to express his deepest gratitude to all those who have played a role in supporting the implementation of this research. Special appreciation is given to the supervising lecturer for his constructive guidance and advice during the research process, as well as to the institution that has provided supporting facilities and infrastructure. Thanks are also extended to family and colleagues who have always provided motivation. The author hopes that the results of this research will make a real contribution to advancing science, particularly in data analysis and information technology.

REFERENCES

[1] P. Studi *et al.,* "Sentiment Analysis of YouTube TvOne Comments on Ustadz Abdul Somad's Deportation from Singapore Using the SVM Algorithm," vol. 13, no. 1, pp. 72–80, 2023, doi: 10.37859/jf.v13i01.4920.

[2] A. A. Chamid, R. Nindyasari, N. Azizah, and A. Hariyadi, “Analysis of public opinion on the governor candidate

debate using LDA and IndoBERT," vol. 4, no. 3, 2025, doi: 10.22219/kinetik.v10i3.2221.
[3] A. Azzam and S. Mahardiningtyas, "Integrating Sentiment Analysis and Quality Function Deployment for Product Development," vol. 7, no. 2, pp. 1–9, 2023, doi: 10.29099/ijair.v7i2.1100.
[4] N. Bayes, C. Dan, and R. Forest, "Sentiment Analysis Of The Sirekap 2024 Playstore Application Post-Presidential Election With A Comparison Of Support Vector Machine (Svm), Naïve Bayes Classifier, And Random Forest Methods," Vol. 11, No. 3, 2025.-Presidential Election With Comparison Of Support Vector Machine (Svm), Naïve Bayes Classifier, And Random Forest Methods," vol. 11, no. 3, pp. 661–670, 2025, doi: 10.25126/jtiik.2025129608.
[5] S. Khomsah, P. Studi, S. Data, J. Tengah, and I. Artikel, "Sentiment Analysis on YouTube Comments Using Word2Vec and Random Forest," vol. 18, no. 1, pp. 61–72, 2021, doi: 10.31515/telematika.v18i1.4493.
[6] K. Di, Y. Tentang, I. Afdhal, R. Kurniawan, I. Iskandar, and R. Salambue, "Application of Random Forest Algorithm for Sentiment Analysis," vol. 5, no. 1, pp. 122–130, 2022, doi: 10.32672/jnkti.v5i1.4004.
[7] M. R. Ramadhan and K. Budiman, "METHOD Sentiment Analysis of Presidential Candidates in 2024: A Comparison of the Performance of Support Vector Machine and Random Forest with N-Gram Method," vol. 3, no. 1, pp. 34–42, 2025, doi: 10.15294/rji.v3i1.8385.
[8] S. A. Penggalih, S. Mujilahwati, and A. A. Bettaliyah, "Twitter Sentiment Analysis to Assess Public Opinion on Jokowi ' s Performance Over Two Periods using the Recurrent Neural Network (RNN) Method," vol. 5, no. 1, 2025, doi: 10.59934/jaiea.v5i1.1736.
[9] X. S. Media, "Sentiment Analysis on Application X on the Use of Red Oil Using the Naïve Bayes Method," vol. 8, no. 1, pp. 149–160, 2025, doi: 10.24014/ijaidm.v8i1.35422.
[10] A. M. Putri, R. Yanuarti, M. Dasuki, and A. M. Susetyo, "Indonesian Journal of Information Systems and Technology Sentiment Analysis of Digital Population Identity Using the Naïve Bayes Multinomial Algorithm Sentiment Analysis of Digital Population Identity Using the Naïve Bayes Multinomial Algorithm data disk," vol. 10, no. 2, pp. 85–95, 2025, doi: 10.32528/justindo.v10i2.2777.
[11] D. Hardiansyah, R. Z. Abdul, and M. Said, "The Classification Method is Used for Sentiment Analysis in My Telkomsel," vol. 8, no. 2, 2024, doi: 10.29099/ijair.v8i2.1229.
[12] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis of Government Policy on Corona Cases Using the Naive Bayes Algorithm," vol. 15, no. 1, pp. 55–64, 2021, doi: 10.22146/ijccs.60718.
[13] V. B. Adiguna, M. Aqqad, J. S. Pt, and H. R. Dp, "Improving Performance Sentiment Analysis Movie Review Film using Random Forest with Feature Selection Information Gain," vol. 8, no. 1, 2024.
[14] R. Sistem, M. Lestandy, A. Abdurrahim, and L. Syafa, "Analysis of COVID-19 Vaccine Tweet Sentiment Using Recurrent," vol. 5, no. 10, pp. 802–808, 2021, doi: 10.29207/resti.v5i4.3308.
[15] R. D. Kurniawan, A. Yohannis, and W. T. Atmojo, "Sentiment Analysis of Getcontact Application Reviews on Google Play Store Using Naive Bayes Algorithm," vol. 6, no. 4, pp. 2848–2858, 2025, doi: 10.52436/1.jutif.2025.6.4.5248.
[16] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," vol. 16, no. 2, pp. 113–124, 2022, doi: 10.22146/ijccs.68903.
[17] A. Muhaddisi, B. N. Prastowo, D. Utami, and K. Putri, "Sentiment Analysis With Sarcasm Detection On Politicians' Instagram," vol. 15, no. 4, pp. 349–358, 2021, doi: 10.22146/ijccs.66375.
[18] N. Anggraini, S. J. Putra, L. K. Wardhani, F. Dhiya, and U. Arif, "A Comparative Analysis of Random Forest, XGboost, and LightGBM Algorithms for Emotion Classification in Reddit Comments," vol. 17, no. 1, pp. 88–97, 2024, doi: 10.15408/jti.v17i1.38651.
[19] H. Herlinda, M. I. Mazdadi, D. Kartini, and I. Budiman, "Implementation of Particle Swarm Optimization on Sentiment Analysis of Cyberbullying using Random Forest," vol. 11, no. 2, pp. 301–309, 2023, doi: 10.30595/juita.v11i2.17920.
[20] C. Ayunda *et al.*, "Comparative Analysis of Machine Learning Algorithms for Sentiment Analysis," vol. 7, no. 2, pp. 80–84, 2022, doi: 10.30591/jpit.v7i2.3547.
[21] A. Mardiah, S. Dillah, D. F. Surianto, N. Fadilah, and S. G. Zain, "Classification of Livin ' by Mandiri customer satisfaction using MLP with BM25 and TF-IDF feature weighting," vol. 4, no. 3, 2025, doi: 10.22219/kinetik.v10i3.2248.
[22] A. N. Syafia, M. F. Hidayattullah, and W. Suteddy, "Comparative Study of SVM and Random Forest Algorithms in Analyzing Sentiment in BTS YouTube Comments," vol. 8, no. 3, pp. 207–212, 2023, doi: 10.30591/jpit.v8i3.5064.

BIBLIOGRAPHY OF AUTHORS

Julianti, a student at the Informatics Engineering Study Program, Faculty of Engineering, West Sulawesi University. Active as a student. This research is a form of final assignment to complete her undergraduate study period.

Farid Wajidi, S.Kom., M.T. is a lecturer in the Informatics Engineering Study Program, Faculty of Engineering, West Sulawesi University. Actively teaches and researches digital image processing, machine learning, and geospatial-based information systems and other implementations of information science.

Musawwir, S.Pd., M.Pd. is a lecturer in the Informatics Engineering Study Program, Faculty of Engineering, West Sulawesi University. Actively teaches and researches educational technology and computer science.