Jurnal Media Teknik Elektro dan Komputer
Vol. 02, No.01 (January - June, 2025)
E-ISSN: 3063-2927

Sentiment Analysis of Instagram Comments on Capital Relocation Using SVM and Random Forest

Zery Mesanda[1], Boy Arnol Sitompul[2]

[1]Universitas Islam Negeri Sumatera Utara; zerymesanda23@gmail.com
[2]Universitas Prima Indonesia; boyarnolsitompul17@gmail.com

ABSTRACT

Social media sentiment analysis has become an important approach in understanding public opinion on strategic issues, including the discourse on the relocation of the national capital. This study aims to compare the performance of Support Vector Machine (SVM) and Random Forest (RF) algorithms in classifying the sentiment of public comments on Instagram. A total of 794 comment data were collected using web scraping techniques with Selenium and BeautifulSoup, then divided into 80% training data and 20% test data. The classification process was conducted after the text preprocessing stage, which included case folding, tokenizing, filtering, and stemming. The experimental results show that SVM achieved an accuracy of 75.0% with precision 0.7200, recall 0.7800, and F1-score 0.7488. Meanwhile, Random Forest performed better with an accuracy of 79.4%, precision of 0.7795, recall of 0.8200, and F1-score of 0.7992. Evaluation based on sentiment class shows that SVM can only achieve a correct rate of 75.0% in the positive class and 75.1% in the negative class, while Random Forest excels with 79.4% in the positive class and 79.3% in the negative class. These findings confirm that Random Forest is more optimal and consistent than SVM in sentiment analysis based on social media comments. This study recommends the use of ensemble learning algorithms such as Random Forest in similar studies, as well as further development with larger datasets and deep learning approaches to improve model accuracy and generalization.

Keywords: Sentiment Analysis, National Capital (IKN), Social Media, Support Vector Machine, Random Forest

Corresponding Author:
Zery Mesanda
Universitas Islam Negeri Sumatera Utara; zerymesanda23@gmail.com

1. INTRODUCTION

The development of information and communication technology has influenced various aspects of society, including public perspectives and participation in strategic national issues. Social media, particularly Instagram, has become one of the most widely used platforms by Indonesians to express opinions, provide comments, and engage in public discussions regarding government policies. With the number of active users continuing to rise, Instagram is not only a platform for sharing images and videos but also a rich source of data on public sentiment toward social and political issues. One of the hotly debated strategic issues in Indonesia is the relocation of the national capital from Jakarta to East Kalimantan. This policy has elicited a wide range of public responses, both supportive and opposed,

which have been expressed massively through social media. With such a large number of comments, manual analysis has become inefficient, necessitating a computational approach to systematically classify public sentiment. This analysis is important because public opinion is a crucial indicator of policy acceptance in the implementation of national development.

The relocation of Indonesia's capital city began to be discussed by the Indonesian government in 2019. Then, through a limited meeting held by the Indonesian government, the President of Indonesia decided to relocate the capital city outside of Java, as stated in the 2020-2024 National Medium-Term Development Plan. The new capital city will be built in the Penajam Paser Utara Regency. During a special committee meeting on the Draft Law on the National Capital (RUU IKN) held in 2022, the Head of Bappenas announced that the new capital in East Kalimantan will be named "Nusantara." The relocation of Indonesia's IKN has naturally sparked various reactions, particularly from the Indonesian public (Saputri & Alita, 2024). The capital city of Nusantara is considered an important focus of research given its strategic role as a center of government, economy, and culture. Sentiment analysis of opinions related to the capital city of Nusantara can provide deep insights into public perceptions and responses to various current issues (Supian et al., 2024). Social media is one of the technologies used by most of the world's population, including in Indonesia, with the number of users reaching around 193 million or around 70% of the total population. This large number of users spans a wide range of ages, from children to the elderly, and reflects the enormous potential of social media in various activities such as advertising, promotion, notifications, and other information distribution. The process of information distribution on social media is two-way, allowing recipients of information to respond through public comment sections, so that everyone can see the responses to a piece of content

When the announcement was made to move the national capital to Nusantara, many Indonesians took to Instagram to voice their opinions on the topic. These opinions, which included both positive and negative comments, had a significant impact on the government. With thousands, if not millions, of Instagram users expressing their views, these comments can be a valuable source of information. However, to effectively utilize this information, accurate analysis is required. One approach to analyzing opinions or sentiments is through sentiment analysis. This study chose to use the Support Vector Machine (SVM) classification algorithm and Random Forest to analyze public sentiment toward the IKN, aiming to obtain more accurate and beneficial results for future decision-making. (Setiawan & Suryono, 2024). In the field of artificial intelligence, sentiment analysis is generally conducted using a machine learning approach. Some popular algorithms that have been used in previous studies are Support Vector Machine (SVM) and Random Forest (RF). SVM is known to be effective in handling labeled data and has high accuracy in text classification, while Random Forest excels in overcoming overfitting and providing better interpretation of influential variables. Previous studies have demonstrated the success of these algorithms in the domains of e-commerce, politics, and public services; however, research specifically analyzing public sentiment regarding the relocation of the IKN through Instagram comments remains limited.

Based on research carried out by (Singh & Tripathi, 2021) sentiment in tweets needs to be classified because it can have both positive and negative impacts. This study utilized a Kaggle dataset containing over 14,000 tweets related to the KFC and McDonald's AI challenge, which was cleaned using the TF-IDF method. Three classification algorithms were tested: SVM, Random Forest, and Decision Tree, with the best results obtained from the Decision Tree algorithm, achieving an accuracy of 88.51%. Evaluation was conducted using accuracy, recall, precision, and F1-score metrics.

In the meantime, research conducted by (Zahoor et al., 2020), related to sentiment analysis aims to classify customer views on products or services through comments on social media. This study analyzed restaurant reviews in Karachi from the SWOT's guide Facebook community, using a dataset containing 4,000 manually annotated records. Comments were categorized as positive or negative and grouped based on taste, atmosphere, service, and value for money. Several classification algorithms were tested, including Naive Bayes, Logistic Regression, SVM, and Random Forest. The best results were obtained from Random Forest with an accuracy of 95%.

Research in this area has been carried out by (Tusar & Islam, 2021) to offer solutions for improving customer satisfaction through Sentiment Analysis based on Natural Language Processing (NLP) and Machine Learning (ML). The study utilized two NLP techniques, namely Bag-of-Words and TF-IDF, as well as several ML classification algorithms, including Support Vector Machine (SVM), Logistic Regression, Multinomial Naive Bayes, and Random Forest, on a large, imbalanced, and multi-class dataset. The results of the study indicate that the best approach was achieved using SVM and Logistic Regression with the Bag-of-Words technique, which achieved an accuracy rate of 77%.

Research conducted by (Aljuaid et al., 2021) using a binary classification approach based on text sentiment analysis, utilizing metadata, number of citations, keywords, and cosine similarity scores between citations. Machine learning models such as SVM, KLR, and Random Forest were evaluated, with the proposed method achieving an F-measure of 0.83 (an improvement of 13.6%) on dataset 1 and 0.67 (an improvement of 8%) on dataset 2 using Random Forest, outperforming current approaches.

In addition, research conducted by (Akter et al., 2021) proposed a binary classification of citations based on sentiment analysis and cosine similarity, which was evaluated using SVM, KLR, and Random Forest. The results show that this method outperforms current approaches with an F-measure of 0.83 (an improvement of 13.6%) on dataset 1 and 0.67 (an improvement of 8%) on dataset 2 using Random Forest.

Based on existing research gaps, this study focuses on analyzing the sentiment of Instagram comments regarding the policy of relocating the national capital (IKN) using Support Vector Machine (SVM) and Random Forest algorithms. The analysis was conducted by collecting comment data from various posts related to the relocation of the IKN that were widely discussed by the public. The data then undergoes preprocessing steps such as text cleaning, normalization, tokenization, and removal of stopwords to prepare it for processing by machine learning models. With this approach, the research aims to provide an in-depth understanding of public perceptions expressed through social media.

The main objectives of this study are: (1) to identify positive and negative public sentiment toward the IKN relocation policy, (2) to evaluate the performance of SVM and Random Forest algorithms in classifying sentiment in comments by comparing accuracy, precision, recall, and f-measure, and (3) to present the distribution of public opinion visually to show trends in public perception. This study is expected to make a significant contribution to the literature on social media-based sentiment analysis, particularly in the context of public policy issues. Additionally, the research findings have practical benefits for the government in understanding the dynamics of public response, thereby providing important input for decision-making processes (Ilhami et al., 2024) and the implementation of national strategic policies such as the relocation of the IKN.

2. LITERATURE REVIEW

2.1. Sentiment Analysis on Social Media

Sentiment analysis is the process of automatically understanding, extracting, and processing text-based information. This process is carried out to obtain implied information about the feelings behind an opinion. It is based on calculations of opinions, sentiments, and feelings (Al Assyam & Hasan, 2023).

Sentiment analysis is also part of the field of text mining, which is often carried out. Text mining is a field of data mining in which structured and unstructured text data is analyzed to obtain valuable information. The text data that has been collected will then be processed, usually used to process and organize existing data by analyzing related information. Text mining is a process of mining information from large amounts of data derived from text. This text mining process is used to process existing text data to obtain the desired information. (Siregar, 2023).

Sentiment analysis on social media is an important field in Natural Language Processing (NLP) that focuses on extracting, identifying, and classifying opinions or subjective expressions contained in text into specific polarities, such as positive, negative, or neutral. With the increasing use of social media, platforms such as Twitter, Facebook, and Instagram have become rich sources of data for describing public opinion. Social media not only provides a large volume of data but also offers the speed and diversity of information that enables analysis to be conducted on a broad scale. Previous studies have utilized social media data to evaluate public sentiment toward commercial products, services, and public policy issues, including strategic government decisions. From a methodological perspective, sentiment analysis on social media is conducted using various approaches, ranging from lexicon-based methods that rely on lists of positive and negative words, to machine learning and deep learning methods that can automatically learn sentiment patterns from data. These approaches provide flexibility and accuracy in understanding public perceptions recorded in digital interactions, making sentiment analysis an important tool in data-driven decision-making.

In addition, sentiment analysis on social media plays a strategic role in the context of public policy, including the issue of relocating the national capital (IKN) in Indonesia. Instagram, as one of the platforms with a high level of interaction, has become a forum for the public to voice their opinions, both in support of and criticism of this policy. The patterns of comments that emerge reflect the collective perceptions of the public, which can be used to measure the level of acceptance, resistance, or concerns among the public. By applying machine learning algorithms such as Support Vector Machine (SVM) and Random Forest, analysis can be conducted more accurately to map the distribution of sentiment. The results of this analysis are not only beneficial for researchers in understanding public opinion but also provide valuable insights for the government in designing communication strategies, follow-up policies, or mitigating social risks that may arise from the implementation of the IKN relocation policy.

2.2. Support Vector Machine

Support Vector Machine (SVM) is one of the most widely used supervised learning algorithms in classification and regression tasks due to its ability to effectively handle high-dimensional data. The basic concept of SVM is to construct an optimal hyperplane that can separate data into different classes with maximum margin. With a larger margin, SVM is expected to improve the model's generalization ability to new data. SVM offers high flexibility through the use of kernel functions, such as linear kernel, polynomial kernel, and radial basis function (RBF), which enable the mapping of non-linear data into

a higher-dimensional space so that it can be separated linearly. This makes SVM effective in handling complex classification problems, including text analysis and sentiment analysis. Previous studies have shown that SVM often outperforms other algorithms in text classification due to its resilience to overfitting on datasets with a large number of features but relatively few samples.

In the context of social media-based sentiment analysis, SVM is widely used to classify public opinion into positive, negative, or neutral categories. The advantage of SVM in utilizing sparse features from text representations such as bag-of-words or TF-IDF makes it one of the algorithms consistently chosen by researchers. Recent studies also confirm that SVM remains competitive compared to ensemble methods and deep learning on small to medium-sized datasets, making it relevant for use in this study to evaluate public sentiment toward policy issues. The advantages of Support Vector Machine (SVM) include its popularity and suitability for classification, as it does not depend on the number of attributes and can address dimensionality issues. Computationally, Support Vector Machine (SVM) can perform training quickly, and its learning techniques can handle challenges related to uncertainty (Septhya et al., 2023).

2.2. Random Forest

Random Forest is one of the bagging-based ensemble learning algorithms developed by Breiman (Irawan et al., 2021). This algorithm works by constructing a number of decision trees on the training data, then making predictions based on the aggregated results (majority voting for classification or averaging for regression). The main advantage of Random Forest is its ability to reduce overfitting, which often occurs in single decision trees, while improving accuracy by leveraging the variation between trees.

Random Forest was developed from the CART (Classification and Regression Trees) method, which is also a method or algorithm of decision tree techniques. What distinguishes the Random Forest method from the CART method is that Random Forest applies the bootstrap aggregating (bagging) method and also random feature selection, which can also be referred to as random feature selection (Suwardika & Suniantara, 2019). Random Forest is a combination of each existing decision tree technique, which are then combined and integrated into a single model. There are three main points in the Random Forest method: the first is performing bootstrap sampling to build prediction trees, each decision tree predicts using random predictors, and then the Random Forest makes predictions by combining the results from each decision tree using majority vote for classification or averaging for regression (Becker et al., 2023).

In the literature, Random Forest is widely used in various text classification tasks, including social media-based sentiment analysis. Researchers have found that Random Forest is capable of handling high-dimensional data derived from text representations such as bag-of-words and TF-IDF with competitive performance compared to other traditional methods. Additionally, Random Forest is relatively robust to imbalanced data because its bootstrap sampling process enhances variability in the learning process.

Several studies have also shown that Random Forest often provides more stable results than margin-based algorithms such as SVM, especially on datasets with limited data but complex features. This makes Random Forest one of the algorithms worth using in social media comment sentiment analysis, including on public policy issues. In the context of this study, Random Forest was evaluated and compared with SVM to determine the more effective model for classifying public opinion on the policy of relocating the National Capital (IKN).

3. METHODS

3.1. Research Design

This section describes the steps involved in applying the Support Vector Machine (SVM) and Random Forest methods for sentiment analysis. This study aims to compare the performance of these two methods in classifying the sentiment of comments related to the development of the Nusantara Capital City (IKN) on Instagram.

The initial stage of this research involves planning to ensure the smoothness and validity of the research process. The Support Vector Machine (SVM) and Random Forest methods will be applied to data collected through comment scraping techniques on several Instagram posts related to the development of the Nusantara Capital City (IKN). This study aims to produce accurate and in-depth sentiment analysis regarding public opinion on the development of the IKN, as well as to obtain a comparison between the two algorithms in terms of accuracy and quality in sentiment classification.

The process began with collecting comments from several Instagram posts discussing IKN Nusantara. Scraping techniques were used to collect comments from Instagram users, which were then saved in an appropriate format, such as CSV or Excel, for further analysis. After the collection period was complete, the comment data was processed and cleaned to ensure data quality and relevance.

Google Colab was used as a platform to implement SVM and Random Forest algorithms in analyzing sentiment data. This process included preparing the working environment in Google Colab, importing the collected data, and applying data preprocessing, such as tokenization, stopword removal, and vectorization using TF-IDF, to prepare the data before analysis. Once the data is ready, both algorithms are implemented to analyze the sentiment of comments and predict public opinion regarding IKN Nusantara. The results of this analysis are then evaluated using metrics such as accuracy, precision, recall, and F1-score. After obtaining the results from the data processing using the three methods, further processing is conducted using a confusion matrix to obtain these evaluation metrics. The results of this evaluation will be used to determine the method with the best performance in sentiment analysis related to the development of IKN Nusantara.

3.2. Model Design

Design is the stage of planning, drawing, and creating sketches or knowledge from several separate elements into a single whole. This results in a flowchart depicting the process flow of the applied model's working mechanism. This model will predict the sentiment of Instagram comments related to the construction of the Nusantara Capital City (IKN) based on patterns found in the training data. The performance of the Support Vector Machine (SVM) and Random Forest models is evaluated using metrics such as accuracy, precision, recall, and F1-score to assess how well these models classify positive and negative sentiments. User comments on Instagram are input into the trained model. The trained SVM and Random Forest models classify the text input according to sentiment categories. The classification results from both models are then displayed, providing users with information about the detected sentiment in comments related to IKN Nusantara. This process concludes with the final results of the sentiment analysis, enabling a more comprehensive assessment of public opinion regarding the development of IKN Nusantara. The research process mechanism is shown in Figure 1.

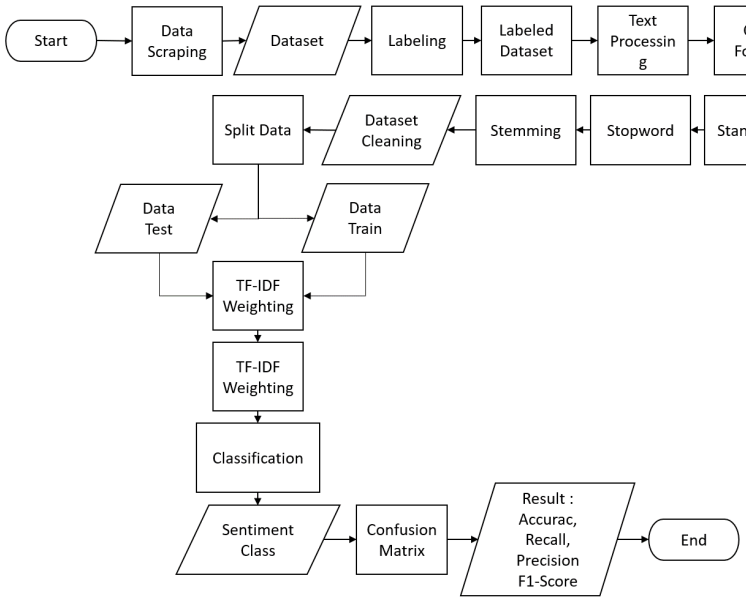

Figure 1. Classification Model Flowchart

The classification stage is the core of the sentiment analysis process carried out in this study. After the text representation is obtained through TF-IDF weighting, the training data is entered into a classification model based on Support Vector Machine (SVM) and Random Forest. SVM is used to find the optimal hyperplane capable of separating positive and negative comments with maximum margin, while Random Forest builds a set of decision trees that are combined through a majority voting mechanism to determine the final class. The selection of these two algorithms is based on their proven reliability in handling high-dimensional text data and their tendency to provide stable performance in social media sentiment classification. The SVM and Random Forest flowcharts, as shown in Figure 1, can be divided into two separate flowcharts for each method. The SVM and Random Forest flowcharts can be viewed sequentially in Figures 2 (Bari Antor et al., 2021).

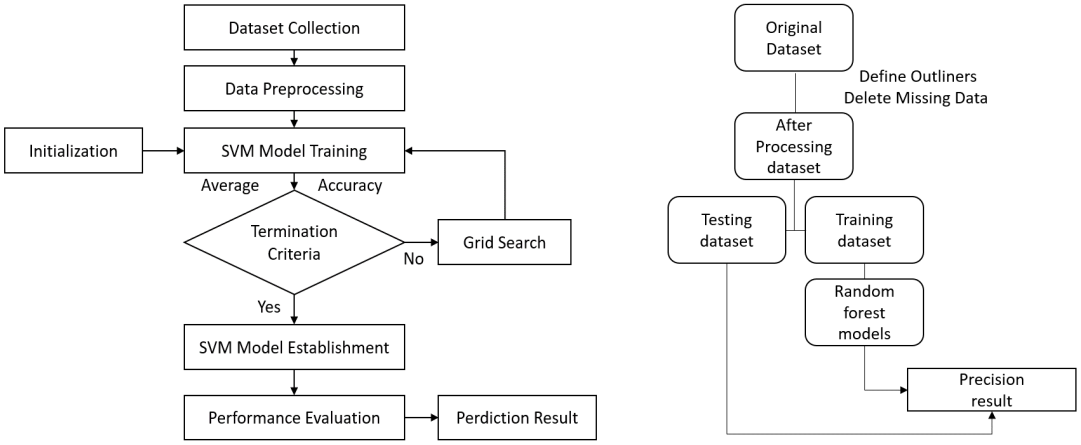

Figure 2. Flowchart Support Vector Machine and Random Forest

3.3. Confusion Matrix

The output of the classification stage is a sentiment class, which is a prediction of the sentiment label on the test data consisting of positive and negative categories. This prediction is then compared

with the actual label using a confusion matrix, which presents the distribution of classification results in the form of true positives, true negatives, false positives, and false negatives. From the confusion matrix, model performance is evaluated using quantitative metrics, namely accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correct predictions from the entire test data, precision indicates the extent to which the model can accurately classify positive/negative comments, recall describes the model's ability to find all relevant comments in each class, while F1-score serves as a measure of the balance between precision and recall. Beberapa uraian rumus untuk menghitung accuracy, precision, recall, and F1-score dapat dijabarkan sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = \frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

Through this stage, the effectiveness of SVM and Random Forest in classifying the sentiment of Instagram comments on the policy of relocating the national capital (IKN) can be analyzed comprehensively. The evaluation results obtained not only provide a quantitative picture of the performance of each algorithm, but also serve as a basis for determining the most appropriate model for capturing public opinion based on social media.

4. RESULTS AND DISCUSSION

In this study, researchers analyzed sentiment toward the development of the Nusantara Capital City (IKN) based on user comments on Instagram. The first stage of this study was to collect comments related to IKN from several posts on Instagram. The data collection process was carried out using a Python library to scrape data.

Scraping is an automated process used to extract specific data from web pages, involving accessing web pages, retrieving the desired content, and storing the data in a more structured format such as CSV or a database. In this study, scraping was performed without relying on a specific time frame, focusing on comments relevant to the Nusantara IKN. The scraping data consists of 794 recent comments related to the development of IKN Nusantara. The scraping process uses the Selenium and BeautifulSoup libraries to extract data from Instagram web pages. After the data was collected, it was saved in .CSV format to make it easier to process using the Python programming language. The collected comment data was then displayed in a table to facilitate further analysis. A sample table of the scraped data is shown in Table 1.

Tabel 1. Instagram Data Scraping Results

Data scraping
Abis upacara pada balik lg ke Jakarta. Pejabat gk mau tinggal disana... Meeeehhhh 😂.
Kawasan yg tidak di restui sebagian besar rakyat Indonesia.

Data scraping
Yang suka buly menangis🤣🤣 yuk tombol like nya yang setuju IKN berjalan dengan baik♡.
Alhamdulillah, perlahan namun pasti insya'allah IKN selesai tepat waktunya ɪᴅ♡ɪᴅ.
Kabarnya warga sekitar gak boleh ikut nonton upaca 17 agustus min??? Selain gitu kabarnya mau sewa alphard & mobil mewah lain 25jt per mobil min??? Kok makin ngadi2 kebijakannya? Mentang2 eksklusif & anyar lalu semena2 gitu??

The preprocessing process aims to clean and prepare the data so that it is ready for further analysis. The data preprocessing process can be described as follows:

4.1. Labeling

At this stage, the sentiment data is labeled. During the labeling stage, each comment obtained from Instagram social media is classified into positive and negative sentiment categories based on the meaning and context of the sentence. This process aims to provide appropriate labels for the data so that it can be used as training data in the classification algorithm. The results of the process and labeling can be shown sequentially in Table 2 and Figure 3.

Tabel 2. Labeled Data

No	Sentimen Input	Sentimen Output
1	Abis upacara pada balik lg ke Jakarta. Pejabat gk mau tinggal disana... Meeeehhhh 😂.	Negatif
2	Kawasan yg tidak di restui sebagian besar rakyat Indonesia.	Negatif
3	Yang suka buly menangis🤣🤣 yuk tombol like nya yang setuju IKN berjalan dengan baik❤️.	Positif
4	Alhamdulillah, perlahan namun pasti insya'allah IKN selesai tepat waktunya ɪᴅ❤️ɪᴅ.	Positif
5	Kabarnya warga sekitar gak boleh ikut nonton upaca 17 agustus min??? Selain gitu kabarnya mau sewa alphard & mobil mewah lain 25jt per mobil??? Kok makin ngadi2 kebijakannya? Mentang2 eksklusif & anyar lalu semena2 gitu??	Negatif

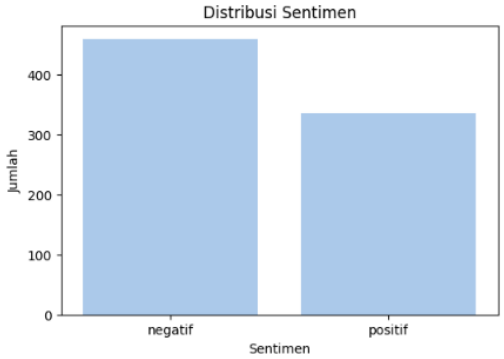

Figure 3. Sentiment Label Visualization

4.2. Text Processing

The Text Processing stage is carried out to convert raw comment data obtained from Instagram into clean text data that is ready for use in the sentiment analysis process. Data taken from social media

usually contains a lot of noise in the form of punctuation marks, emoticons, abbreviations, excessive capital letters, and irrelevant words. Therefore, a series of pre-processing stages are carried out to make the text representation more structured. The process begins with case folding, which converts all letters to lowercase so that differences in capitalization do not affect the analysis. Next, cleansing is performed to remove irrelevant characters such as numbers, punctuation marks, emoticons, symbols, hyperlinks, and other special characters. After that, the tokenization stage is used to separate sentences into word fragments (tokens) so that they can be processed further. The next stage is stopword removal, which is the removal of common words that do not contribute significantly to the meaning. The text processing process can be described as follows:

1. Case Folding
   During the case folding stage, all capital letters are converted to lowercase, but numbers, punctuation marks, emoticons, and symbols are left unchanged as they will be cleaned up in the next stage (cleansing).
2. Filtering
   After going through the case folding stage, text that previously contained capital letters has been converted entirely to lowercase letters. However, the result still contains characters that are irrelevant for analysis, such as punctuation marks, numbers, emoticons, and certain symbols that do not contribute meaningfully to the formation of text features. Therefore, the filtering stage is carried out by removing all non-alphabetic characters, including emoticons, punctuation marks, and numbers, leaving only words that represent the main meaning of the comments.
3. Tokenizing
   The tokenizing stage is the process of breaking down the filtered text into the smallest units in the form of words or tokens. This process is very important because sentiment analysis and machine learning-based text modeling require individual word representations in order to be recognized as features. At this stage, sentences that are still strings of words are separated based on spaces, resulting in a list of words that are easier to process further, such as stemming and stopword removal.
4. Stopword Removal (Stangword)
   The stopword removal stage is the process of removing common words that are considered to have no significant meaning in the analysis. These words usually appear very often in sentences, but do not contribute significantly to the sentiment context or the core meaning of the text. By removing stopwords, the data becomes more concise, relevant, and focused on keywords that really influence the further analysis process.
5. Stemming
   The stemming process is the next stage after stopword removal, in which every word relevant to sentiment analysis is converted to its root form. The purpose of stemming is to ensure that words with varying forms are still recognized as a single feature by the SVM or Random Forest methods. The results of the text processing can be seen in Table 3.

Tabel 3. Text Processing Results

No	Input (Stopword Removal)	Output (Stemming)
1	[abis, upacara, balik, jakarta, pejabat, tinggal, meeeeehhh]	[habis, upacara, balik, jakarta, pejabat, tinggal, meeeeehhh]
2	[kawasan, restui, sebagian, besar, rakyat, indonesia]	[kawasan, restu, sebagian, besar, rakyat, indonesia]
3	[suka, buly, menangis, like, setuju, ikn, berjalan, baik]	[suka, buly, tangis, like, setuju, ikn, jalan, baik]
4	[alhamdulillah, perlahan, pasti, insyaallah, ikn, selesai, tepat, waktunya]	[alhamdulillah, perlahan, pasti, insyaallah, ikn, selesai, tepat, waktu]
5	[kabarnya, warga, nonton, upaca, agustus, sewa, alphard, mobil, mewah, makin, ngadi, kebijakannya, eksklusif, anyar, semena, gitu]	[kabar, warga, tonton, upaca, agustus, sewa, alphard, mobil, mewah, makin, jadi, kebijakan, eksklusif, anyar, semena, gitu]

4.3. Splitting Data

After a series of data cleaning and processing steps have been completed, the next step is to split the data. This is done to separate the dataset into two main parts: training data and testing data(Gibran et al., 2024) (Rifki et al., 2025). Training data is used to train the model to recognize patterns and relationships within the data (Rambe et al., 2025). Meanwhile, testing data is used to evaluate the model's performance after the training process is complete. The data division scheme is implemented with an 80:20 composition, with each dataset allocated 635 training data points (80%) and 159 testing data points (20%). This is done to avoid overfitting, a situation where the model performs very well on training data but fails to produce good results on previously unseen data. The process continues by applying

4.4. Classification Model

The text classification modeling process uses a Pipeline approach in machine learning. This Pipeline consists of two main components, namely the TF-IDF Vectorizer. This stage performs text feature extraction using the Term Frequency–Inverse Document Frequency (TF-IDF) method. The goal is to convert raw text into a numerical representation in the form of a word weight matrix. TF-IDF weights emphasize important words in a specific document that are rare across the entire corpus, thereby improving the quality of feature representation for classification. The text classification model architecture using SVM and Random Forest can be seen in Figure 4.

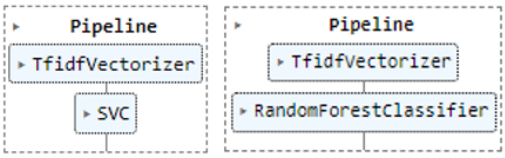

Figure 4. Classification Modeling Flow SVM and Random Forest

Furthermore, the results of sentiment classification analysis are visualized in the form of a word cloud to provide an overview of the frequency and dominance of words that appear in public opinion regarding the relocation of the national capital (IKN). This visualization is used to identify positive and

negative sentiment trends based on the words most frequently used in the analyzed text. The negative and positive visualizations can be seen in Figures 4 and 5, respectively.

Figure 5. Positive Sentiment Word Cloud

Figure 6. Negative Sentiment Word Cloud

4.5. Model Testing

After conducting training using the Random Forest algorithm, the model is then tested by printing a classification report and displaying a matrix, resulting in sentiment classification results in the form of sentiment labels. The classification labels generated are compared with the actual labels to determine the accuracy, precision, recall, and F1-score of the model used on the dataset. After training using the Support Vector Machine algorithm, the model is tested by generating a classification report and displaying a matrix, resulting in sentiment classification outcomes in the form of sentiment labels. The generated classification labels are compared with the actual labels to determine the accuracy, precision, recall, and F1-score values of the model used on the dataset. The results of the model testing can be seen in Figures 7 and 8.

	precision	recall	f1-score	support
0	0.7822	0.7226	0.7512	1640
1	0.7200	0.7800	0.7488	1500
accuracy			0.7500	3140
macro avg	0.7511	0.7513	0.7500	3140
weighted avg	0.7525	0.7500	0.7500	3140

Figure 7. SVM Model Testing Results

	precision	recall	f1-score	support
0	0.8101	0.7680	0.7885	1000
1	0.7795	0.8200	0.7992	1000
accuracy			0.7940	2000
macro avg	0.7948	0.7940	0.7939	2000
weighted avg	0.7948	0.7940	0.7939	2000

Figure 8. Random Forest Model Testing Results

4.6. Method Performance

Comparison of the performance of two classification methods, namely Support Vector Machine (SVM) and Random Forest, in classifying comments into two categories: positive and negative. The bar chart shows the distribution of correct classification results, with the number of positive and negative comments identified by each method. SVM successfully classified 258 positive comments with an accuracy rate of 75.0% and 338 negative comments with an accuracy rate of 75.1%. Meanwhile, Random Forest demonstrated better performance with 273 positive comments (79.4%) and 357 negative comments (79.3%). The results of the comparison of comment classification results using SVM and Random Forest can be seen in Figure 9.

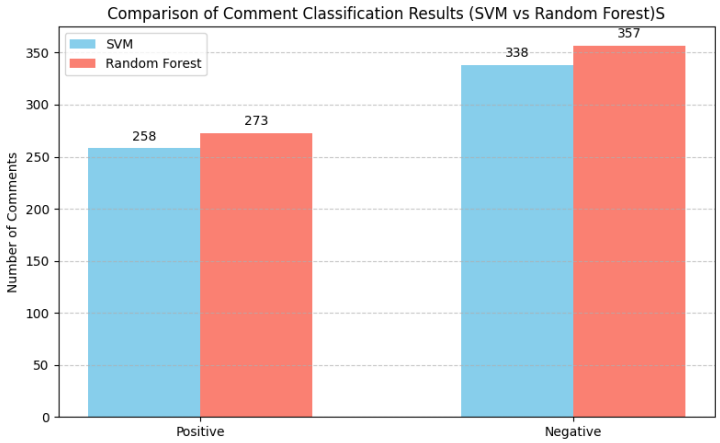

Figure 9. Comparison of Comment Classification Results (SVM vs Random Forest)

Meanwhile, calculation details including the number of errors (false classifications) and their percentages were also analyzed. SVM produced 86 errors in the positive class and 112 errors in the negative class, while Random Forest produced 71 errors in the positive class and 93 errors in the negative class. These findings confirm that Random Forest has a higher accuracy rate and a lower number of classification errors compared to SVM, both in the positive and negative classes. This indicates that Random Forest is more reliable in handling comment data with a total of 794 data points, making it a more effective method for sentiment classification in the context of this study. The tabulation of the comparison of comment classification results using SVM and Random Forest can be seen in Table 4.

Tabel 4. Method Performance Percentages (SVM vs Random Forest)

Methods	Class	Number of True	Percentation of True (%)	Number of False	Percentation of False (%)	Total
SVM	Positive	258	75,0%	86	25,0%	344
SVM	Negative	338	75,1%	112	24,9%	450
Random Forest	Positif	273	79,4%	71	20,6%	344
Random Forest	Negatif	357	79,3%	93	20,7%	450

REFERENCES

Akter, M. T., Begum, M., & Mustafa, R. (2021). Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors. *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 40–44. https://doi.org/10.1109/ICICT4SD50815.2021.9396910

Al Assyam, H. D., & Hasan, F. N. (2023). Analisis sentimen Twitter terhadap perpindahan ibu kota negara ke IKN nusantara menggunakan orange data mining. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(1), 341–349.

https://doi.org/https://doi.org/10.30865/klik.v4i1.957

Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., & Tanvir Afzal, M. (2021). Important citation identification using sentiment analysis of in-text citations. *Telematics and Informatics*, *56*, 101492. https://doi.org/https://doi.org/10.1016/j.tele.2020.101492

Bari Antor, M., Jamil, A. H. M. S., Mamtaz, M., Monirujjaman Khan, M., Aljahdali, S., Kaur, M., Singh, P., & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict Alzheimer's disease. *Journal of Healthcare Engineering*, *2021*(1), 9917919.

Becker, T., Rousseau, A.-J., Geubbelmans, M., Burzykowski, T., & Valkenborg, D. (2023). Decision trees and random forests. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(6), 894–897. https://doi.org/10.1016/j.ajodo.2023.09.011

Gibran, M. K., Rifki, M. I., Hasugian, A. H., Siahaan, A. T. A. A., Sahputra, A., & Ong, R. (2024). Sentiment Analysis of Platform X Users on Starlink Using Naive Bayes. *Instal: Jurnal Komputer*, *16*(03), 210–220. https://doi.org/10.54209/jurnalinstall.v16i03.240

Ilhami, A. M., Hadiansyah, M. N. H., Baihaqi, A. A., & Khalid, I. P. (2024). Priority Decision Making System for Educational Fund Assistance Letters Using Top-Down Parsing Method. *Jurnal Media Teknik Elektro Dan Komputer*, *01*(01), 19–26.

Irawan, D., Perkasa, E. B., Yurindra, Y., Wahyuningsih, D., & Helmud, E. (2021). Perbandingan Klassifikasi SMS Berbasis Support Vector Machine, Naive Bayes Classifier, Random Forest dan Bagging Classifier. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *10*(3), 432–437. https://doi.org/10.32736/sisfokom.v10i3.1302

Rambe, M. R. A., Zufria, I., & Rifki, M. I. (2025). Analisis Sentimen Masyarakat pada Platform Media Sosial X (Twitter) terhadap Pelantikan Kabinet Merah Putih Menggunakan Bernoulli Naïve Bayes. *DEVICE: JOURNAL OF INFORMATION SYSTEM, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY*, *6*(1), 83–102. https://doi.org/10.46576/device.v6i1.6360

Rifki, M. I., Gibran, M. K., Hasugian, A. H., & Solihin, M. D. (2025). PEMODELAN DAN EVALUASI PREDIKSI RSRP MENGGUNAKAN ARTIFICIAL NEURAL NETWORK UNTUK OPTIMASI KUALITAS LAYANAN JARINGAN KOMUNIKASI NIRKABEL. *INFORMATIKA*, *17*(1), 392–401. https://doi.org/10.36723/juri.v17i1.751

Saputri, G. A., & Alita, D. (2024). Analisis Sentimen Twitter Terhadap Pemindahan Ibu Kota Negara Menggunakan Support Vector Machine. *Jurnal Informatika: Jurnal Pengembangan IT*, *9*(3), 213–223.

Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Rahmaddeni, R., Irawan, Y., & Hayami, R. (2023). Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru: Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(1), 15–19. https://doi.org/https://doi.org/10.57152/malcom.v3i1.591

Setiawan, A., & Suryono, R. R. (2024). Analisis Sentimen Ibu Kota Nusantara menggunakan Algoritma Support Vector Machine dan Naïve Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *8*(1), 183–192. https://doi.org/https://doi.org/10.29408/edumatic.v8i1.25667

Singh, J., & Tripathi, P. (2021). Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm. *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, 193–198. https://doi.org/10.1109/CSNT51715.2021.9509679

Siregar, A. M. (2023). Analisis Sentimen Pindah Ibu Kota Negara (IKN) Baru pada Twitter Menggunakan Algoritma Naive Bayes dan Support Vector Machine (SVM). *Faktor Exacta*, *16*(3). https://doi.org/http://dx.doi.org/10.30998/faktorexacta.v16i3.16703

Supian, A., Revaldo, B. T., Marhadi, N., Efrizoni, L., & Rahmaddeni, R. (2024). Perbandingan Kinerja Naïve Bayes Dan Svm Pada Analisis Sentimen Twitter Ibukota Nusantara. *Jurnal Ilmiah Informatika*, *12*(01), 15–21. https://doi.org/https://doi.org/10.33884/jif.v12i01.8721

Suwardika, G., & Suniantara, I. K. P. (2019). Analisis Random Forest Pada Klasifikasi CART Ketidaktepatan Waktu Kelulusan Mahasiswa Universitas Terbuka. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *13*(3), 179–186. https://doi.org/10.30598/barekengvol13iss3pp177-184ar910

Tusar, M. T. H. K., & Islam, M. T. (2021). A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data. *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, 1–4. https://doi.org/10.1109/ICECIT54077.2021.9641336

Zahoor, K., Bawany, N. Z., & Hamid, S. (2020). Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning. *2020 21st International Arab Conference on Information Technology (ACIT)*, 1–6. https://doi.org/10.1109/ACIT50332.2020.9300098