Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 575-592
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4230

Analyzing Public Sentiment on the Relocation of Indonesia's Capital to Kalimantan as the Ibu Kota Nusantara Using Logistic Regression

Agisni Zahra Latifa[1], Warih Maharani*[2]

[1]Data Science, School of Computing, Telkom University, Indonesia
[2]Informatics, School of Computing, Telkom University, Indonesia

Email: [2]wmaharani@telkomuniversity.ac.id

Received : Dec 17, 2024; Revised : Jan 13, 2025; Accepted : Jan 13, 2025; Published : Apr 26, 2025

Abstract

The Ibu Kota Nusantara (IKN) relocation project aims to equalize economic development and reduce the burden on Jakarta, but has elicited mixed reactions from the public, including both support and opposition. Therefore, this study applies machine learning-based sentiment analysis, using Logistic Regression to explore public opinion on the relocation, and leveraging social media data from platform X to gain insights into information, opinions, and public reactions. The Textblob, VADER, and SentiWordNet labeling methods employ a majority vote of the three labels to determine the final label. In order to achieve data balance, SMOTE is employed in this study. Moreover, this study applies a combination of preprocessing, N-gram, and TF-IDF to illuminate the impact of this combination on model performance. The results indicate that the combination of preprocessing Scenario 3 with unigram, bigram, trigram, and TF-IDF feature extraction yields the best performance, achieving a precision of 0.7641, recall of 0.7767, F1-score of 0.7634, and accuracy of 0.7641. This research demonstrates the efficacy of proper preprocessing and feature extraction in enhancing the performance of the Logistic Regression model for sentiment classification, thereby contributing to the analysis of public opinion on IKN policy regarding other issues in the future.

Keywords : *IKN Sentiment, Logistic Regression, N-gram, Preprocessing, Public Opinion, TF-IDF.*

1. INTRODUCTION

The relocation of the Indonesian capital from Jakarta to Kalimantan as the Ibu Kota Nusantara (IKN) is carried out to accelerate Indonesia's economic equality, reduce development inequality, and create jobs [1]. The move is also expected to ease the burden on Jakarta, including welfare, health, and environmental issues such as congestion, air pollution, flooding, clean water shortages, crime, and land subsidence[2] [3]. The relocation plan has been under discussed the era of since Indonesia's first president Soekarno's era and continued until President Susilo Bambang Yudhoyono's era, and finally received serious attention under the leadership of President Joko Widodo's [4]. The process of relocating the capital also involves several stages of good planning, such as feasibility studies, infrastructure planning, public consultation, and stakeholder involvement [5]. While the relocation of the capital city was well intentioned, it still triggered many public responses and discussions on support and opposition, as the project was underway. This indicates the necessity of comprehending public sentiment, as it exerts a substantial influence on the public's perspective regarding the decision to relocate the capital city[6]. With this understanding, the government can gain valuable insights to predict public views, craft more appropriate responses, prevent potential conflicts, and make wiser decisions in the future [7]. Therefore, this research is conducted to analyze the public sentiment related to the policy of relocating Indonesia's capital policy using sentiment analysis based on machine learning.

In the digital era, social media has become the primary means of communication for individuals to express their views and opinions [8][9]. One such platform is X, formerly known as Twitter. X is a social media that allows its users to write and express opinions or actions through tweets [10][11]. The hashtag and keyword features help users find the latest information on trending topics [12][13]. This research will therefore utilise tweet data on the X platform to explore information, understand opinions, and public reactions to IKN policies.

The background of the problem, to find out and analyze tweets related to IKN, sentiment analysis research is carried out. Sentiment analysis is a technique in Natural Language Processing (NLP) that employs machine learning to identify and categorize emotions, opinions, or perspectives in text as negative [14][15][16]. This research uses Logistic Regression (LR), a common machine learning method for text classification and sentiment analysis [17]. Logistic Regression is a classification algorithm used for sentiment analysis and produces positive or negative classes [18][19]. Logistic Regression produces The most effective sentiment prediction evaluates accuracy, recall, precision, and F1-score using the N-Gram and TF-IDF methods [20]. The classification approach that results in consistent test results is Logistic Regression [21]. In addition, Logistic Regression provides class probabilities that can help in strengthening sentiment analysis results [22][23].

There are several studies that have conducted sentiment analysis using the Logistic Regression method. Previous research [17] shows that Logistic Regression with TF-IDF using VADER labeling has an accuracy of 85.22%, superior to KNN Textblob 73.64% in sentiment analysis of the Covid-19 vaccine. The study found that the highest accuracy of the three classification models, LR, SVM and NNB, was 86.54% for bigram [23]. Research by Ramadhon et al. [24] sentiment analysis of capital city relocation using N-grams with bigram and K-Nearest Neighbor reached accuracy of 66%. Twitter data on capital city relocation used TF-IDF and three algorithms [25]. The findings revealed that the SVM algorithm exhibited the highest accuracy, at 97.72%, in comparison to Logistic Regression and KNN. Sinha et al. [26] discussed sentiment analysis regarding the conflict between Russia and Ukraine by comparing methods between KNN, Decision Tree and Logistic Regression with the results obtained accuracy of 88.89%, 90.45%, and 94.58%. Research by Wahyuningsih et al. [27] compared LR, NB, and random forest algorithms to predict students' reasoning using data sources from Kaggle, LR has the highest accuracy of 94.34%.

Previous research has discussed sentiment analysis, but no one has specifically focused on the topic of IKN using the Logistic Regression method. Previous studies focused on comparisons between methods or on a single type of simple feature extraction. This study differs from previous studies because this study uses logistic regression by applying three labeling techniques and testing different combinations of data preprocessing and feature extraction methods. This combination is done to explore and identify the most effective approach for generating sentiment classification using Logistic Regression related to IKN topics.

This research is aims to analyze the sentiment of the public over the capital relocation policy of the Indonesian government through the application of machine learning. The present study utilizes a machine learning algorithm, specifically Logistic Regression. During the analysis process, a combination of preprocessing and feature extraction is employed to ascertain the most effective approach for conducting sentiment analysis of public opinion concerning the relocation of the National Capital. It is from this approach that this research is expected to contribute to an understanding of the perception and dynamics of public opinion towards the policy

2. METHOD

The following figure illustrates the research process for sentiment classification of tweets regarding the IKN topic on social media X, which outlines the stages from data collection to model evaluation. The initial stage of the process involves the collection of data from X. The data then undergoes preprocessing, where the data is already labeled using TextBlob, VADER, and SentiWordNet. The next step is to divide the data into training and test data, followed by feature extraction using a combination of N-gram and TF-IDF. If the training data is imbalanced, SMOTE is used. The process ends with the creation and evaluation of the model.

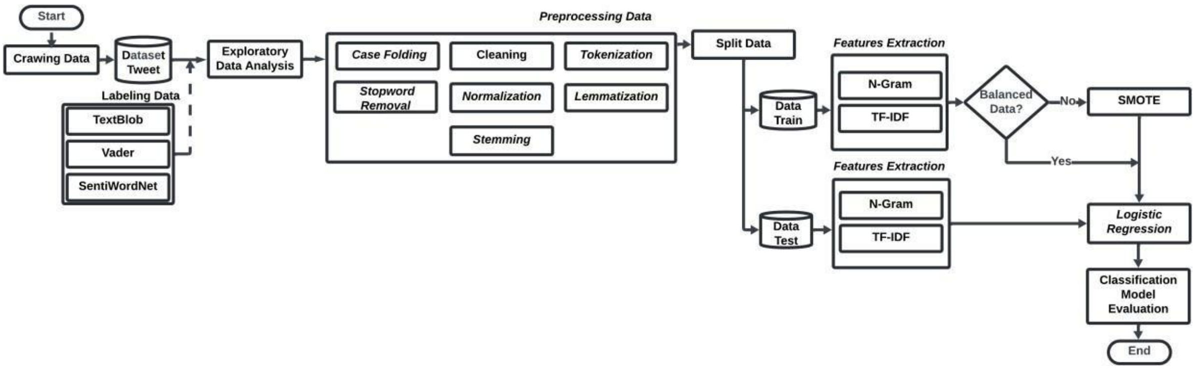

Figure 1. Research Workflow

2.1. Data Collection

The data employed in this study were derived from social media tweets on the X platform. The data was extracted using APIs and tokens on X, employing the Python programming language, with the keywords “IKN”, “Ibu Kota Baru”, and “Ibu Kota Nusantara”. A total of 7,077 rows of data were collected, starting in February 2024, after Jakarta lost its status as Indonesia's state capital on February 15, and continuing until September 2024. However, the data collection was not conducted continuously throughout each month, resulting in an uneven distribution of collection periods.

2.2. Data Labeling

In this process, tweet labeling is conducted by three annotators, using Textblob, VADER, and SentiWordNet. The final label is determined by a majority vote of the three annotators.

2.2.1. Textblob

TextBlob is a Python module that facilitates text analysis and utilizes the Natural Language Toolkit (NLTK) to calculate sentiment scores[17], [28]. An averaging technique is applied to compute a popularity score for the entire text. The resulting score is expressed on a scale of -1 to 1, the value -1 is indicative of negative sentiment, whereas +1 is indicative of positive sentiment. The following is equation (1) for the TextBlob labeling process [29].

$$Label = \{ \begin{matrix} Positive \quad if\ 0 < score \leq 1 \\ Negative \quad if\ -1 \leq score < 0 \\ Neutral\ if\ score == 0 \end{matrix} \qquad (1)$$

2.2.2. VADER

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a sentiment analysis that employs a synthesis of many sentiment lexicons, these are classified by semantic orientation, which can be positive or negative [17]. VADER identifies the emotional tone of text written in everyday language, making it a valuable resource for social media texts, reviews, and online comments [28]. The following is equation (2) for the VADER labeling process [29].

$$Label = \{ \begin{matrix} Positive \quad if\ score \geq 0.05 \\ Negative \ \ if\ score \leq -0.05 \\ Neutral\ if\ \ -0.05 < score < 0.05 \end{matrix} \tag{2}$$

2.2.3. SentiWordNet

SentiWordNet is a lexicon used for automatic analysis of sentiment in textual data. SentiWordNet creates synsets (sets of synonyms) [28]. The synset is associated with a positive and negative popularity score. A synset represents a set of words with similar meanings. The following is equation for the SentiWordNet labeling process [30].

$$synset_{score} = positive_{score} - negative_{score} \tag{3}$$

$$\frac{\sum synset_{score}}{\sum Tokens} \tag{4}$$

$$overall_{Score} > 0 : "postive" \tag{5}$$

$$overall_{Score} < 0 : "negative" \tag{6}$$

2.3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an important process for overview, study pattern, and identify trends from the data from the data used in the early stages of data analysis [31]. The purpose of EDA is to understand the data before using machine learning [31]. Furthermore, EDA also provides insight into the quality and characteristics of the data [31].

2.4. Preprocessing Data

The process of data preprocessing entails the cleansing of tweet text to prepare it for subsequent use and processing. The objective is to enable the utilization of the data by other models or algorithms. The preprocessing stage is of particular importance in the context of sentiment analysis, particularly in the case of social media data, where users often provide informal, unstructured opinions that may be affected by typos and other forms of noise [32]. This research involved several scenarios with different types of preprocessing steps applied [33].

a. Case Folding is a method of converting characters in a string to uppercase or lowercase. It is employed to transform the entire text into lowercase.
b. Cleaning is performed to preprocess the text data by removing all non alphabetic characters from the tweet, such as numbers, symbols, punctuation marks, and URLs.
c. Tokenisasi is the process of parsing a sentence into its constituent words.
d. Stopword Removal is a process of removes unimportant words from tweets.
e. Normalization is the process of standardizing the text by correcting spelling variations and ensuring consistency in word forms.
f. Lemmatization is the process of transforming words into their fundamental form.
g. Stemming is a method that removes words modified by the insertion of affixes.

2.5. Data Splitting

This study use data partitioning to verify the model. The dataset is divided into two types, 80% for training data 20% for and testing data. Training data is employed for the purpose of learning and predicting labels, while testing data is used to assess the performance of the model following the learning process [34].

2.6. N – Gram

In this study, the data were initially divided and subsequently processed using N-gram. An N-gram is defined as an n-word subsequence of a text or string [35]. This study utilizes combination N-grams. The purpose of this N-gram application is to comprehend the context of the data by analyzing the sequence of words [36]. After the N-gram process, the results will be represented numerically with TF-IDF.

2.7. TF - IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used to assess the importance of a word in a document or sentence [37][38]. TF-IDF is employed for the purpose of assigning weights to individual words, wherein the weight of a given word is determined by the frequency with which it appears in a given document. TF assesses the significance of frequently occurring terms inside a text, while IDF evaluates the prevalence of a certain word across several documents [39][40]. The formula for TF-IDF[20].

1. The TF value calculation is based on the formula (7).

$$TF = \frac{\mathrm{n_{t,j}}}{\text{Document word total}} \tag{7}$$

$\mathrm{n_{t,j}}$ = Frequency of term (t) in document (j)

2. The IDF value calculation is based on formula (8).

$$IDF = log(\frac{D}{DF_w} + 1) \tag{8}$$

D = The overall count of documents in the dataset.
DF_w = The count of documents where the word (w) appears.

3. TF-IDF value calculation is based on formula (9).

$$(TF - IDF)_{t,j} = TF_t(D_j) \times IDF_t \tag{9}$$

2.8. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is used when a dataset is uneven or imbalanced. This method offers several advantages, including its ability to address issues related to data imbalance, minimize prediction errors, and prevent machine learning overfitting models [41]. SMOTE is an oversampling technique to address the sample data for the minority class [42]. The formula SMOTE [43], Y' represents synthetic data generated for the minority class, where Y^i is a sample from the minority class, Y^j is a randomly selected k-nearest neighbor of Y^i, and γ is a random value between 0 and 1.

$$Y' = Y^i + (Y^j - Y^i) * \gamma \tag{10}$$

2.9. Logistic Regression

Logistic Regression uses a sigmoid function to generate probabilities [17]. Variables in the form of input that represents the feature vector $[x_1, x_2, \ldots\ldots, x_n]$ and the output of the classification is either 1 or 0. When the feature represents the number of words in a document, $P(y = 1 \mid x)$ is probability of the document having a positive sentiment, and $P(y = 0 \mid x)$ is probability of the document having a negative sentiment is represented by the value. The following is the formula for Logistic Regression [22].

$$z = (\sum_{i=1}^{n} w_i x_i) + \mathrm{b} \tag{11}$$

In equation (11), x_i multiplied by each weight w_i plus a bias term b where the weight w_i represents how important the input feature is to the positive or negative classification decision. There is a similar formulation to equation (11) which is equation (12).

$$z = \mathrm{w} . \mathrm{x} + \mathrm{b} \tag{12}$$

To create a probability, input z through the sigmoid function (13).

$$f(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

2.10. Confusion Matrix

Confusion matrix evaluates a model's performance on test data by comparing its predictions with the actual values [25]. Confusion matrix has 4 part values which can be seen in Table 1 [17].

Table 1. Confusion Matrix

	Predicted	
Actual	**Positive**	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

This research uses accuracy, precision, recall, and F1-score. Accuracy measures the model's overall ability to recognise data [44]. Precision helps ensure the positive predictions the model makes are correct [44]. Recall focuses on how much positive data was successfully recognized [44]. Especially when the data is unbalanced, the F1-score to provide a balanced evaluation. Formula for calculating the performance of the metrics [45].

1. Accuracy of a model is determined by its efficacy in classifying data as either positive or negative. The accuracy calculation is calculated based on the formula (14).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{14}$$

2. Precision is a calculation used to determine the effectiveness of the model in identifying positive label predictions. The precision calculation is calculated based on the formula (15).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{15}$$

3. Recall is a calculation that determines the percentage of positive cases correctly identified as positive. The recall calculation is based on the formula (16).

$$\text{Recall} = \frac{TP}{TP+FN} \tag{16}$$

4. F1-Score is a calculation used to determine the average comparison of precision and recall. The calculation of f1- score is calculated based on formula (17).

$$\text{F1} - \text{Score} = 2 \times \frac{(Precision\ x\ Recall)}{(Precision + Recall)} \tag{17}$$

3. RESULT

3.1. Data Collection and Data Labeling

The dataset comprised 7,077 tweets collected from February 2024 to November 2024, following the relocation of IKN. This research employs Indonesian tweet data containing the keywords "IKN," "Ibu Kota Baru," and "Ibu Kota Nusantra”. The final label is the most frequent of the three sentiment labels from TextBlob, VADER, and SentiWordNet. The following example illustrates the data labeling procedure, and the initial data samples in Table 2.

Table 2. Example Dataset

Teks	Annotator			Label
	Textblob	VADER	SentiWordNet	
@EllyKoro Kl masy ada dayak marah slesai dah tu IKN bakal jd hambalang jilid 2 dan jd kota hantu selamanya	Negative	Negative	Positive	Negative
@meraaahputiiih Progres IKN menciptakan momentum positif untuk pertumbuhan ekonomi dan pembangunan.	Positive	Positive	Positive	Positive
IKN bukan hanya ibu kota baru tapi simbol pemerataan ekonomi untuk Indonesia	Positive	Negative	Negative	Negative
@Simanjunta9Nico @m1n4_95 Indonesia malah bikin ibu kota baru....	Positive	Negative	Positive	Positive
IKN dongkrak ekonomi kalimantan	Negative	Positive	Negative	Negative

Table 2 displays a sample dataset annotated with TextBlob, VADER, and SentiWordNet. The final label is determined based on the most votes from the three methods. However, as shown in Table 2, the labeling results do not fully align with the context or the actual meaning of the tweets. Specifically, there are tweets that should be labeled positive but are labeled negative, and vice versa. This indicates that the method used still has weaknesses, even though it involves three annotators.

3.2. Exploratory Data Analysis

The next stage of the process was the EDA, which is conducted with the aim of understanding the tweet data. The EDA conducted in this study includes distribution of sentiment, the frequency of words, the distribution of tweet length, and the visualization of the data.

3.2.1. Sentimen Distribution

The Sentiment Distribution stage is conducted to ascertain whether the distribution of positive or negative sentiment categories is balanced or imbalanced within the dataset. The results of the sentiment distribution are presented in Table 3.

Table 3. Sentiment Data Distribution

Sentiment	Total	Percentage
Positive	3,661	52%
Negatie	3,416	48%

As shown in Table 3, following the official decision on 15 February 2024, the majority of people expressed support for the relocation and development of the IKN. A total of 52% of respondents expressed support for the relocation and development of IKN, while 48% indicated opposition or disagreement. These findings provide insight into the public's perceptions of the Indonesian government's policies.

3.2.2. Word Frequency

This word frequency step is to calculate the frequency of words in order to identify trends and patterns of words that appear frequently in tweets pertaining to IKN topics. The results of the word frequency analysis are presented in Table 4 for data with the keywords "IKN" and "Ibu Kota Nusantara", and in Table 5 for data with the keyword "Ibu Kota Baru".

Table 4. Word Frequncy with IKN

No	Word	Frequency
1.	ikn	5,466
2	di	2,620
3	dan	1,501
4	yg	1,460
5	yang	1,002
6	ke	972
7	itu	886
8	kota	803
9	ada	789
10	ini	682

Table 5. Word Frequency with Ibu Kota Baru

No	Word	Frequency
1.	kota	1,374
2	ibu	1,314
3	baru	1,159
4	di	676
5	ikn	434
6	dan	418
7	yang	388
8	nusantara	286
9	yg	251
10	akan	221

Table 4 and Table 5 demonstrate that the most frequently occurring word for the keyword "IKN" is "ikn," with a frequency of 5,466. As for "Ibu Kota Baru", the most frequent words are "kota" and "ibu" with a frequency of 1,374 and 1,314 respectively. Additionally, there are numerous other words, such as "di", "ke", and similar terms, that are dominant in both tables. These words lack specific meanings but are utilized to link sentences and organize opinions in tweets. This table is beneficial for elucidating word usage patterns in tweets pertaining to the subject of IKN.

3.2.3. Distribution of Tweet Lengths

The tweet length distribution stage aims to measure the length of tweets related to the IKN topic, so as to provide an understanding of how people express their opinions. For results, the tweet length distribution is shown in Figure 2 and Figure 3.

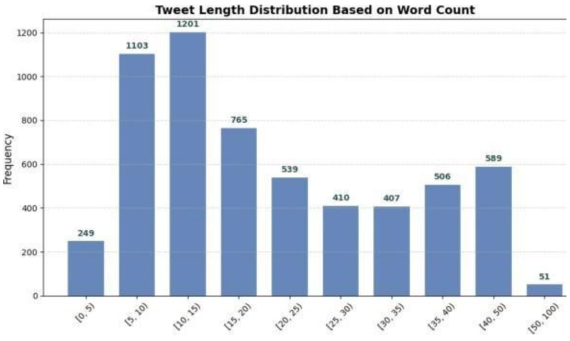

Figure 2. Distribution Tweet with IKN

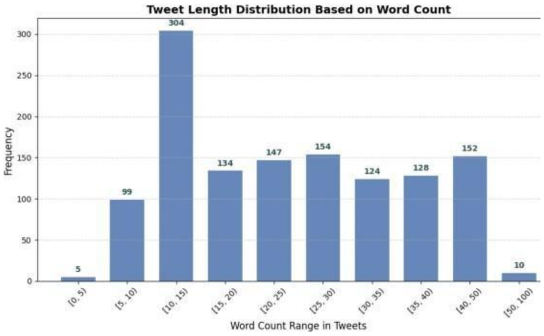

Figure 3. Distribution Tweet with Ibu Kota Baru

As shown in Figures 2 and 3, the distribution of tweet length with the keyword "IKN" displays a tendency for users to express opinions on the platform in compact tweets, the highest frequency is observed in the range of 10–15 words, with 1,201 tweets, followed by the range of 5–10 words, with 1,103 tweet. Similarly, the distribution of tweet length for the keyword “ibu kota baru" exhibits the same pattern, with the highest frequency in the 10-15 word range with 304 tweets, though the total frequency is lower. Overall, the keyword "IKN" is more frequently used than ″Ibu Kota Baru″ to express opinions, and the dataset contains tweets with varying lengths, ranging from short to long.

3.2.4. Word Clouds

The word cloud stage was performed to visualize the most frequently occurring words related to the IKN topic. However, words that already exist in Tables 4 and 5 are not displayed in the word cloud. Word cloud results in Figure 4 and Figure 5.

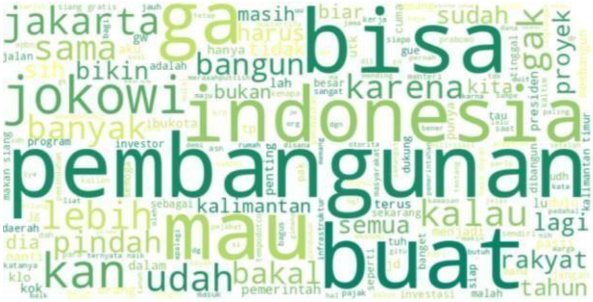

Figure 4. Word Clouds Sentiment Positive

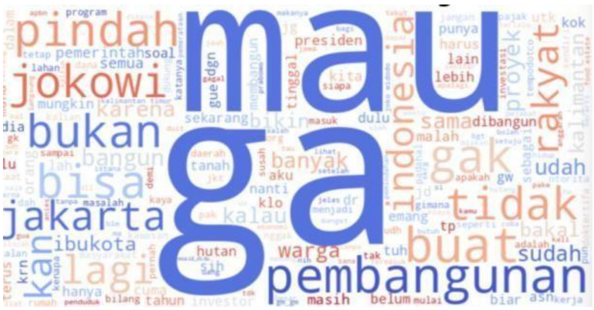

Figure 5. Word Clouds Sentiment Negative

Based on the word cloud in Figures 4 and 5, the frequent appearance of words such as "pembangunan", "bisa", "buat", and "mau" in positive sentiment indicates the public's hope for the development of IKN, especially in terms of bringing progress and benefits to Indonesia. In contrast, the negative sentiments are characterised by the frequent appearance of words such as "ga", "tidak", "pindah", "karena", and "proyek" which reflect disapproval and doubt about the capital move plan. The use of the words "ga", "tidak", and "gak" indicates a rejection, whereas terms such as "bukan" and "malah" underscore the view that this project is not seen as the primary solution to the main issue. This highlights a divergence in opinion, where positive sentiments are predominantly driven by hope and aspirations while negative sentiments are more in criticism and concern.

3.3. Preprocessing Data

The preprocessing performed in this study consists of several types. The sample data displayed after preprocessing is shown in Table 6. The preprocessing process ensures data is fit for purpose. This study employed five preprocessing scenarios:

1. Scenario 1 : Case Folding, Cleaning, Tokenization.
2. Scenario 2 : Case Folding, Cleaning, Tokenization, Stopword Removal.
3. Scenario 3 : Case Folding, Cleaning, Tokenization, Normalization, Lemmatization.
4. Scenario 4 : Case Folding, Cleaning, Tokenization, Stopword Removal, Stemming
5. Scenario 5 : Case Folding, Cleaning, Tokenization, Stopword Removal, Normalization, Lemmatization.

Table 6. Sample Preprocessing Text

Preprocessing	Before	After
Scenario 1	@meraaahputiiih Progres IKN menciptakan momentum positif untuk pertumbuhan ekonomi dan pembangunan.	progres ikn menciptakan momentum positif untuk pertumbuhan ekonomi dan pembangunan
Scenario 2	@meraaahputiiih Progres IKN menciptakan momentum positif untuk pertumbuhan ekonomi dan pembangunan.	progres ikn menciptakan momentum positif untuk pertumbuhan ekonomi pembangunan
Scenario 3	@meraaahputiiih Progres IKN menciptakan momentum positif untuk pertumbuhan ekonomi dan pembangunan.	progres ibu kota nusantara cipta momentum positif untuk tumbuh ekonomi dan bangun
Scenario 4	@meraaahputiiih Progres IKN menciptakan momentum positif untuk pertumbuhan ekonomi dan pembangunan.	progres ikn cipta momentum positif untuk tumbuh ekonomi bangun
Scenario 5	@meraaahputiiih Progres IKN menciptakan momentum positif untuk pertumbuhan ekonomi dan pembangunan.	progres ibu kota nusantara cipta momentum positif untuk tumbuh ekonomi bangun

3.4. Data Splitting

The tweet dataset is split 2 parts, 80% training and 20% test. This division is done because the training data is used to train the model so that the model can recognize patterns from the data, while the test data is used to measure the performance of the model that has been made based on the training data. The results of the data separation are shown in Table 7.

Table 7. Data Splitting Reults

Data	Positive	Negative	Total
Training Data	2,925	2,736	5,661
Testing Data	736	680	1,416

3.5. Feature Extraction

This feature extraction stage follows preprocessing and uses training and test data. Each word in the tweet is weighted with N-grams and TF-IDF to capture information about adjacent word pairs. In this study, 25 experiments were conducted with five preprocessing scenarios and five feature extractions N-grams and TF-IDF. The following combinations of feature extraction were tested :

a. Combination 1 : Unigram + TF- IDF
b. Combination 2 : Bigram + TF-IDF
c. Combination 3 : Trigram + TF-IDF
d. Combination 4 : Unigram + Bigram + TF-IDF
e. Combination 5 : Unigram + Bigram + Trigram + TF-IDF

3.5.1. N – Gram

After obtaining the data from the preprocessing stage, the next step is to apply N-grams to detect adjacent word pairs. The sample results of the N-gram process of Scenario 1 are shown in Table 8.

Table 8. N – Gram Results

N - Gram	Word
Unigram	'ikn', 'memajukan', 'bangsa'
Bigram	'ikn memajukan', 'memajukan bangsa'
Trigram	'ikn memajukan bangsa'
Unigram + Bigram	'ikn', 'memajukan', 'bangsa', 'ikn memajukan', 'memajukan bangsa'
Unigram+Bigram+Trigram	'ikn', 'memajukan', 'bangsa', 'ikn memajukan', 'memajukan bangsa', 'ikn memajukan bangsa'

3.6. SMOTE

The data that is already in numerical form is subjected to SMOTE for training data after the N-gram and TF-IDF feature extraction phases. This is done to balance the data, as the data in this study is not balanced. The findings of SMOTE are presented in Table 9.

Table 9. After SMOTE Results

Training Data	Before SMOTE	After SMOTE
Positive	2,925	2,925
Negative	2,736	2,925

3.7. Logistic Regression Model Performance

The present study aims to evaluate the performance of the Logistic Regression algorithm using various combinations of feature extraction using sentiment classification. A confusion matrix evaluates a variety of scenarios, incorporating precision, recall, accuracy and the F1-score. This experiment aims to find out which Scenario and feature extraction works best in the classification process. Table 10 shows the results of implementing Scenarios 1 to 5 with N-gram feature extraction and TF-IDF.

Table 10. Evaluasion Results

Preprocessing Scenario	Combination Feature Extraction	Precision	Recall	F1-Score	Accuracy
Scenario 1 (Case Folding, Cleaning, Tokenization)	**Unigram + TF-IDF**	**0.7433**	**0.7434**	**0.7433**	**0.7436**
	Bigram + TF-IDF	0.6389	0.6391	0.6389	0.6391
	Trigram + TF-IDF	0.6339	0.6156	0.5969	0.6081
	Unigram + Bigram + TF-IDF	0.7329	0.7318	0.7321	0.7331
	Unigram + Bigram + Trigram + TF - IDF	0.7368	0.7351	0.7354	0.7366
Scenario 2 (Case Folding, Cleaning, Tokenization, Stopword Removal)	Unigram + TF-IDF	0.7159	0.7162	0.7159	0.7161
	Bigram + TF-IDF	0.6378	0.6365	0.6338	0.6342
	Trigram + TF-IDF	0.6236	0.5847	0.5422	0.5734
	Unigram + Bigram + TF-IDF	**0.7219**	**0.7223**	**0.7217**	**0.7218**
	Unigram + Bigram + Trigram + TF - IDF	0.7182	0.7186	0.7181	0.7182
Scenario 3 (Case Folding, Cleaning, Tokenization, Normalization, Lemmatization)	Unigram + TF-IDF	0.7425	0.7425	0.7425	0.7429
	Bigram + TF - IDF	0.6836	0.6838	0.6835	0.6836
	Trigram + TF - IDF	0.6328	0.6326	0.6313	0.6314
	Unigram + Bigram + TF - IDF	0.7619	0.7610	0.7613	0.7620
	Unigram + Bigram + Trigram + TF - IDF	**0.7641**	**0.7767**	**0.7634**	**0.7641**
Scenario 4 (Case Folding, Cleaning, Tokenization, Stopword Removal, Stemming)	**Unigram + TF - IDF**	**0.7329**	**0.7333**	**0.7329**	**0.7331**
	Bigram + TF - IDF	0.6396	0.6392	0.6376	0.6377
	Trigram + TF - IDF	0.6326	0.5941	0.5556	0.5833
	Unigram + Bigram + TF - IDF	0.7241	0.7240	0.7240	0.7246
	Unigram + Bigram + Trigram + TF - IDF	0.7220	0.7219	0.7220	0.7225
Scenario 5 (Case Folding, Cleaning, Tokenization, Stopword Removal, Normalization, Lemmatization)	Unigram + TF - IDF	0.7331	0.7334	0.7329	0.7331
	Bigram + TF - IDF	0.6719	0.6719	0.6709	0.6709
	Trigram + TF - IDF	0.6415	0.6412	0.6398	0.6398
	Unigram + Bigram + TF - IDF	**0.7391**	**0.7394**	**0.7387**	**0.7387**
	Unigram + Bigram + Trigram + TF - IDF	0.7294	0.7298	0.7294	0.7295

As illustrated in Table 10, the performance of the sentiment classification model is influenced by the preprocessing stage and the combination of feature extraction methods employed. Scenario 3, which incorporates unigram, bigram, trigram, and TF-IDF, yields the highest values for precision of 0.7641, recall of 0.7767, F1-score of 0.7634 and accuracy of 0.7641. Although stopword removal was not implemented during the preprocessing phase, scenario 3 still had the highest performance. The findings of this research indicate that the absence of stopword removal yields superior outcomes in some cases compared to its use. This is evident from the comparison of Scenario 1 with Scenario 2, and Scenario 3 with Scenario 5. This finding indicates that stopword removal does not always improve model performance, because in some cases stopwords may contain important information [46].

Furthermore, Scenario 3, which extends Scenario 1, shows a significant improvement after incorporating normalization and lemmatization processes. While Scenario 1, which relied on basic preprocessing steps such as case folding, cleaning, and tokenization, performed reasonably well, the addition of normalization and lemmatization in Scenario 3 led to a substantial boost in accuracy. This improvement is due to the fact that normalization and lemmatization play a key role in text analysis by simplifying the text and reducing word variation, ultimately enhancing the model's ability to recognize patterns more easily [47].

The combination of unigram + bigram + trigram + TF-IDF has proven to yield optimal results, depending on the preprocessing scenario applied. This will definitely show the potential of this combination in the capturing of patterns and context in tweets about the IKN topic. Single word unigram, two consecutive word bigram, and three consecutive word trigram help the model understand meaning or patterns within tweets. In contrast, bigram + TF-IDF and trigram + TF-IDF showed the worst performances for both cases of preprocessing, which may indicate that bigram and trigram solely are somewhat poor at capturing the pattern in tweets. But combined with unigram, bigram, and trigram could be much better in capturing the meaning and patterns in tweets, complex variation of words, and understanding of the relation between words. The finding signifies that the very best preprocessing including normalization and lemmatization along with feature selection is highly vital to improving the performance of the Logistic Regression model, for sentiment analysis in IKN related topics, during this work.

Figure 6 shows a visualization of the model evaluation results for the different combinations of preprocessing and feature extraction, with a comparison of model performance based on precision, recall, F1 score, and accuracy metrics. This graph provides a clear picture of the performance of each combination and shows the combination that produces the best performance.

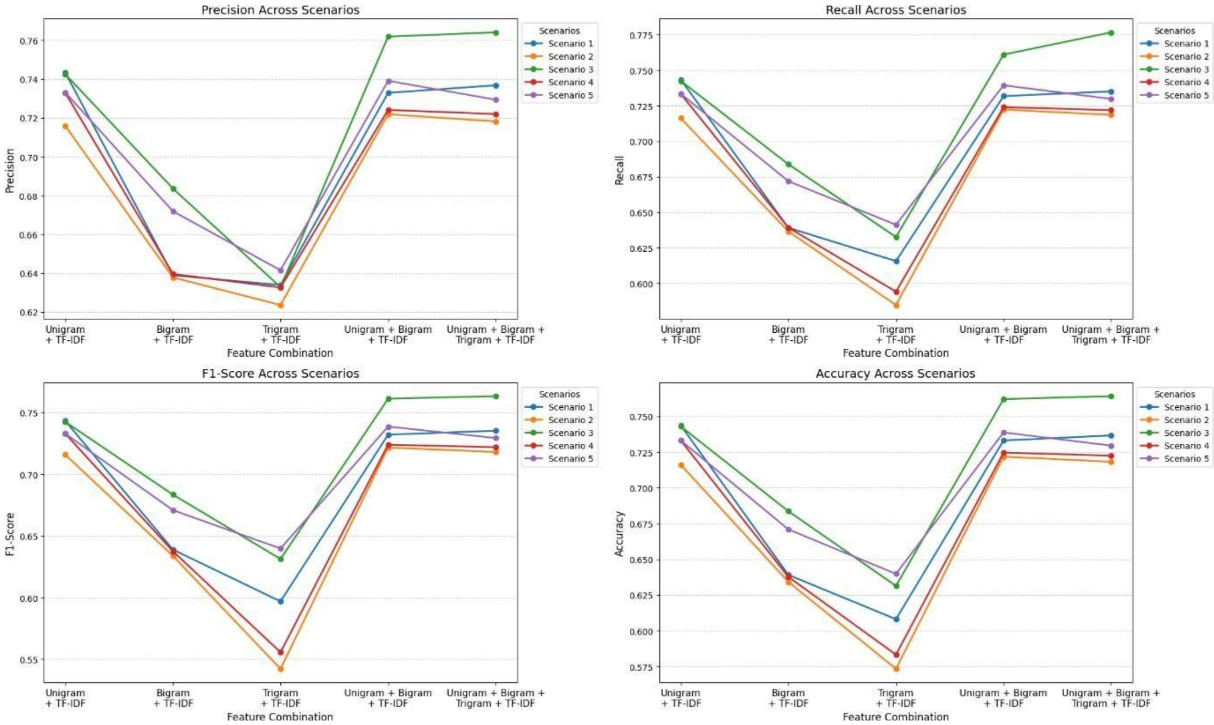

Figure 6. Comparative Analysis of Preprocessing Scenarios and Feature Extraction Combinations

The graphs generally demonstrate that the performance of the model is contingent upon the combination of features and scenarios employed. More complex feature combinations, such as unigram + bigram + trigram + TF-IDF, exhibit more stable and superior performance across all metrics in comparison to simple combinations. However, simple feature combinations, such as unigram + TF-IDF, also yielded optimal results. Conversely, the bigram + TF-IDF and trigram + TF-IDF feature combinations exhibited a consistent decline in performance across various metrics, suggesting that the effectiveness of bigram and trigram features is diminished.Scenario 3 demonstrated a notable advantage, consistently achieving the best results across different feature combinations, particularly in complex scenarios.

4. DISCUSSIONS

There are several previous studies in Table 11 that focus on the same topic as this research, namely the relocation of the capital, conducted by Ramadhon et al. [24], Andi et al. [48], Nurharjadmo et al. [6], dan Muliawan et al. [49].

Table 11. Comparison of Studies on Capital Relocation

Author	Model Algorithm	Feature Extraction	Result
Ramadhon et al.	K-Nearest Neighbor	N-Gram	66% with bigram
Andi et al.	CNN	Word2Vec	70.3%
Nurharjadmo et al.	SVM, Decision Tree	TF-IDF	85.78% with SVM
Muliawan et al.	Naïve Bayes, K-Nearest Neighbor, Random Forest.	-	65.26% with NB

This study highlights significant differences compared to previous research, such as [24], [48], [6], and [49] particularly in the choice of model algorithms and the techniques employed for feature extraction. In addition, this study employs a dataset from 2024. The labeling technique employs three annotators and a combination of preprocessing scenarios and N-gram + TF-IDF feature extraction, achieving an accuracy of 0.7641. This result is higher than Ramadhon et al., Muliawan et al., and Andi et al., but lower than Nurharjadmo et al.

Results of the study are illustrated in Figure 6, while Table 10 indicates that Scenario 3 with the unigram + bigram + trigram + TF-IDF, which is higher than other scenarios and feature extraction combinations. The addition of normalisation and lemmatisation, excluding stop word removal, significantly enhanced the model's performance in this study. Normalisation standardises text by rectifying spelling inaccuracies, hence facilitating processing and analysi [47]. Lemmatization converts words to their basic form, reduces data size, and captures the core meaning of the word[47]. This study aligns with [50] [51] [52] [53], The use of normalization and lemmatization to improve model performance in text analysis. However, removing stopwords can sometimes lead to a reduction in classification accuracy, articularly in documents or texts that depend on prepositions, conjunctions, and auxiliary verbs to provide context or meaning[46]. This phenomenon can be attributed to the fact that Figures 4, 5, and Tables 4 and 5 present the results of word frequency analysis, frequently employing conjunctions, adverbs, extensions, and pronouns.Consequently, the use of stop words can play a crucial role in preserving sentence structure, significant information, and semantic content.For instance, the words "tidak" and "ga" in this particular sentiment offer crucial insights into the content of the tweet.

The combination of unigram + bigram + trigram features improved the performance of the model in this study. The results are in aligns with [54][55], This combination of extraction is able to understand sentence patterns and the complexity of opinion tweets. This technique is also effective in capturing relationships between words in tweets and can even identify emotional phrases such as “IKN keren” and “IKN jelek”. In addition, based on Figures 2 and 3 in the distribution of tweet lengths, this topic has tweets of various lengths, so the combination of unigram + bigram + trigram is effective in helping the model capture patterns and relationships between words, both in simple and complex tweets. The use of TF-IDF gives appropriate weight to the word patterns in the tweets. The experiments conducted generally demonstrate that a suitable combination of preprocessing and feature extraction techniques can enhance the performance of a model in sentiment classification in Logistic Regression, particularly in specific topic datasets such as IKN. Previous studies have indicated that the selection of appropriate preprocessing and feature extraction techniques can lead to improvements in the performance of a model [56], [57].

This research provides an implementation of the NLP method in performing sentiment analysis of public opinion on IKN. Various scenarios of preprocessing and feature extraction combinations are applied to improve the performance of the most effective Logistic Regression model. Furthermore, it provides a tangible solution for comprehending public sentiment surrounding ssues, such as the IKN matter, while concurrently functioning as a reference point for the future analysis of policy or issues in terms of public opinion.

5. CONCLUSION

This study employs a Logistic Regression model to classify sentiments pertaining to the relocation of the capital city to the IKN.The research demonstrates that Logistic Regression can classify sentiment in positive and negative classes, with the selection of the right preprocessing and feature extraction processes that can help the model improve model performance.The results demonstrate that the combination of Scenario 3 process and unigram + bigram + trigram + TF-IDF feature extraction is most effective. Normalization and lemmatization assist in reducing word variation.With unigram + bigram + trigram + TF-IDF feature extraction and without stopword removal, the model is able to capture tweet patterns and relationships between words in tweets. This approach shows potential for understanding public sentiment on strategic issues, which can support data driven decision making.

This research demonstrates that Logistic Regression is an effective method for classifying sentiment regarding the relocation of the capital to IKN. The practical implications of sentiment classification are highly relevant for a range of stakeholders, including the government in understanding public opinion, journalists in analysing emerging issues, and companies in assessing public sentiment about the IKN project. Furthermore, the study contributes to the development of sentiment analysis systems that can be applied to other issues in the future, such as evaluating public policies or analysing opinions on social media. Future research should focus on using more and larger diverse datasets to improve model generalisation. Exploration of deep learning based models, such as transformers, could offer valuable insights by comparing their performance with the current method. The importance of selecting appropriate preprocessing and feature extraction techniques is also highlighted, tailored to the characteristics of the data, in order to significantly improve the performance of sentiment classification models in future applications.

REFERENCES

[1] D. A. Permatasari and . S., “New National Capital City (IKN) in Legal Polemic,” *KnE Social Sciences*, pp. 862–871, Jan. 2024, doi: 10.18502/kss.v8i21.14801.

[2] T. Baharuddin, A. Nurmandi, Z. Qodir, H. Jubba, and M. Syamsurrijal, “Bibliometric Analysis of Socio-Political Research on Capital Relocation: Examining Contributions to the Case of Indonesia,” *Journal of Local Government Issues*, vol. 5, no. 1, pp. 17–31, Mar. 2022, doi: 10.22219/logos.v5i1.19468.

[3] R. Bonita and D. Wadley, “Disposal of government offices in Jakarta pending relocation of the Indonesian capital: an application of multi-criteria analysis,” *Property Management*, vol. 40, no. 4, pp. 591–628, Jul. 2022, doi: 10.1108/PM-10-2020-0068.

[4] A. Hadinata and B. M Cynthia, “The state capital relocation policy and pandemic covid-19: a literature review,” *Journal of Economics and Business Letters*, vol. 1, pp. 24–27, 2021, doi: 10.32479/ijefi.11348.

[5] M. Nurdin and T. Baharuddin, “Capacity Building Challenges and Strategies in the Development of New Capital City of Indonesia,” *Jurnal Bina Praja*, vol. 15, no. 2, pp. 221–232, Aug. 2023, doi: 10.21787/jbp.15.2023.221-232.

[6] W. Nurharjadmo, F. Ansoriyah, and M. A. Khadija, “Analyzing Public Perception using Aspect Based Sentiment Analysis: Case Study of Capital Relocation Planning of Indonesia,” in

2024 7th International Conference on Informatics and Computational Sciences (ICICoS), 2024, pp. 191–196. doi: 10.1109/ICICoS62600.2024.10636903.

[7] T. Xie, Y. yao Wei, W. fan Chen, and H. nan Huang, “Parallel evolution and response decision method for public sentiment based on system dynamics,” *Eur J Oper Res*, vol. 287, no. 3, pp. 1131–1148, Dec. 2020, doi: 10.1016/j.ejor.2020.05.025.

[8] N. Helberger, “The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power,” *Digital Journalism*, vol. 8, no. 6, pp. 842–854, Jul. 2020, doi: 10.1080/21670811.2020.1773888.

[9] H. Ihsaniyati, S. Sarwoprasodjo, P. Muljono, and D. Gandasari, “The Use of Social Media for Development Communication and Social Change: A Review,” Feb. 01, 2023, *MDPI*. doi: 10.3390/su15032283.

[10] A. Sesagiri Raamkumar, M. Erdt, H. Vijayakumar, E. Rasmussen, and Y.-L. Theng, “Understanding the Twitter Usage of Humanities and Social Sciences Academic Journals,” *Proceedings of the Association for Information Science and Technology*, vol. 55, pp. 430–439, 2018, doi: 10.1002/pra2.2018.14505501047.

[11] A. Kumar and A. Jaiswal, “Systematic literature review of sentiment analysis on Twitter using soft computing techniques,” *Concurr Comput*, vol. 32, p. e5107, Nov. 2019, doi: 10.1002/cpe.5107.

[12] G. Karuna, P. Anvesh, C. S. Singh, K. R. Reddy, P. K. Shah, and S. S. Shankar, “Feasible Sentiment Analysis of Real Time Twitter Data,” in *E3S Web of Conferences*, Oct. 2023, p. 10. doi: 10.1051/e3sconf/202343001045.

[13] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, “An ‘Infodemic’: Leveraging high-volume twitter data to understand early public sentiment for the Coronavirus disease 2019 outbreak,” *Open Forum Infect Dis*, vol. 7, no. 7, Jul. 2020, doi: 10.1093/ofid/ofaa258.

[14] M. P. Abraham, “Feature Based Sentiment Analysis of Mobile Product Reviews using Machine Learning Techniques,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 2289–2296, 2020, doi: 10.30534/ijatcse/2020/210922020.

[15] S. Sonare and M. Kamble, “Sentiment Analysis in polished Product Based Inspections data using existing supervised machine learning approach,” in *2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society, TRIBES 2021*, 2021, pp. 1–6. doi: 10.1109/TRIBES52498.2021.9751663.

[16] F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, and S. Qiao, “Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 9, pp. 4332–4345, 2022, doi: 10.1109/TNNLS.2021.3056664.

[17] F. Fazrin, O. N. Pratiwi, and R. Andreswati, “Perbandingan Algoritma K-Nearest Neighbor dan Logistic Regression pada Analisis Sentimen terhadap Vaksinasi Covid-19 pada Media Sosial Twitter dengan Pelabelan Vader Dan Textblob,” *eProceedings of Engineering*, vol. 10, no. 2, p. 1596, Apr. 2023.

[18] P. Harjule, A. Gurjar, H. Seth, and P. Thakur, “Text Classification on Twitter Data,” in *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, 2020, pp. 160–164. doi: 10.1109/ICETCE48199.2020.9091774.

[19] O. Shobayo, S. Adeyemi-Longe, O. Popoola, and B. Ogunleye, “Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4 and Logistic Regression: A Data-Driven Approach,” *Big Data and Cognitive Computing*, vol. 8, no. 11, Nov. 2024, doi: 10.3390/bdcc8110143.

[20] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The impact of features extraction on the sentiment analysis,” *Procedia Comput Sci*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[21] H. R. Alhakiem and E. B. Setiawan, “Aspect-Based Sentiment Analysis on Twitter Using Logistic Regression with FastText Feature Expansion,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 840–846, 2022, doi: 10.29207/resti.v6i5.4429.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. 2024.

[23] A. Poormina and S. P. K, "A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques," *2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)* , pp. 493–496, 2020, doi: 10.1109/ICACCS48705.2020.9074312.

[24] M. I. Ramadhon, A. Arini, F. Mintarsih, and I. M. M. Matin, "N-Gram and K-Nearest Neighbor Algorithm for Sentiment Analysis on Capital Relocation," in *2021 9th International Conference on Cyber and IT Service Management, CITSM 2021*, IEEE, 2021. doi: 10.1109/CITSM52892.2021.9587919.

[25] E. Sutoyo and A. Almaarif, "Twitter sentiment analysis of the relocation of Indonesia's capital city," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1620–1630, Aug. 2020, doi: 10.11591/eei.v9i4.2352.

[26] A. Sinha, B. Rout, S. Mohanty, S. R. Mishra, H. Mohapatra, and S. Dey, "Exploring Sentiments in the Russia-Ukraine Conflict: A Comparative Analysis of KNN, Decision Tree and Logistic Regression Machine Learning Classifiers," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1068–1076. doi: 10.1016/j.procs.2024.04.101.

[27] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 349–356. doi: 10.1016/j.procs.2024.03.014.

[28] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach," *Soc Netw Anal Min*, vol. 13, no. 1, Dec. 2023, doi: 10.1007/s13278-023-01030-x.

[29] S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," *IEEE Access*, vol. 11, pp. 11811–11826, 2023, doi: 10.1109/ACCESS.2023.3242290.

[30] N. M. Sham and A. Mohamed, "Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches," *Sustainability (Switzerland)*, vol. 14, no. 8, p. 4723, Apr. 2022, doi: 10.3390/su14084723.

[31] V. Da Poian *et al.*, "Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry," *Frontiers in Astronomy and Space Sciences*, vol. 10, May 2023, doi: 10.3389/fspas.2023.1134141.

[32] S. Sumayah, F. Sembiring, and W. Jatmiko, "ANALYSIS OF SENTIMENT OF INDONESIAN COMMUNITY ON METAVERSE USING SUPPORT VECTOR MACHINE ALGORITHM," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 1, pp. 143–150, 2023, doi: 10.52436/1.jutif.2023.4.1.417.

[33] H. T. Duong and T. A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput Soc Netw*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40649-020-00080-x.

[34] A. Larasati, R. P. Editya, Y. W. Chen, and V. E. B. Darmawan, "The Effect of Data Splitting Ratio and Vectorizer Method on the Accuracy of the Support Vector Machine and Naïve Bayes Model to Perform Sentiment Analysis," in *ICEEIE 2023 - International Conference on Electrical, Electronics and Information Engineering*, Institute of Electrical and Electronics Engineers Inc., 2023, p. 1. doi: 10.1109/ICEEIE59078.2023.10334755.

[35] K. U. Wijaya and E. B. Setiawan, "Hate Speech Detection Using Convolutional Neural Network and Gated Recurrent Unit with FastText Feature Expansion on Twitter," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 619–631, Sep. 2023, doi: 10.26555/jiteki.v9i3.26532.

[36] C. C. Wang, M. Y. Day, and C. L. Wu, "Political Hate Speech Detection and Lexicon Building: A Study in Taiwan," *IEEE Access*, vol. 10, pp. 44337–44346, 2022, doi: 10.1109/ACCESS.2022.3160712.

[37] S. Fahmi, L. Purnamawati, G. F. Shidik, M. Muljono, and A. Z. Fanani, "Sentiment analysis of student review in learning management system based on sastrawi stemmer and SVM-PSO," in *Proceedings - 2020 International Seminar on Application for Technology of Information and*

Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020, IEEE, 2020, pp. 643–648. doi: 10.1109/iSemantic50169.2020.9234291.

[38] G. A. Dalaorao, A. M. Sison, and R. P. Medina, "Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy," in *TSSA 2019 - 13th International Conference on Telecommunication Systems, Services, and Applications, Proceedings*, 2019, pp. 282–85. doi: 10.1109/TSSA48701.2019.8985458.

[39] H. Aljuaid, R. Iftikhar, S. Ahmad, M. Asif, and M. Tanvir Afzal, "Important citation identification using sentiment analysis of in-text citations," *Telematics and Informatics*, vol. 56, 2021, doi: 10.1016/j.tele.2020.101492.

[40] U. Bhattacharjee, P. K. Srijith, and M. S. Desarkar, "Term Specific TF-IDF Boosting for Detection of Rumours in Social Networks," in *2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019*, IEEE, 2019. doi: 10.1109/COMSNETS.2019.8711427.

[41] M. Umer *et al.*, "Scientific papers citation analysis using textual features and SMOTE resampling techniques," *Pattern Recognit Lett*, vol. 150, pp. 250–257, Oct. 2021, doi: 10.1016/j.patrec.2021.07.009.

[42] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed, and S. Hussain, "Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis," *IEEE Access*, vol. 9, pp. 84296–84305, 2021, doi: 10.1109/ACCESS.2021.3087827.

[43] H. Hairani, T. Widiyaningtyas, and D. Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV International Journal on Informatics Visualization*, vol. 8, pp. 1310–1318, Dec. 2024, doi: 10.62527/joiv.8.3.2283.

[44] S. Yang and G. Berdine, "Confusion matrix," *The Southwest Respiratory and Critical Care Chronicles*, vol. 12, no. 53, pp. 75–79, Oct. 2024, doi: 10.12746/swrccc.v12i53.1391.

[45] F. A. Ramadhan and P. H. Gunawan, "Sentiment Analysis of 2024 Presidential Candidates in Indonesia: Statistical Descriptive and Logistic Regression Approach," *2023 International Conference on Data Science and Its Applications, ICoDSA 2023*, pp. 327–332, 2023, doi: 10.1109/ICODSA58501.2023.10276417.

[46] M. Işik and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," 2020, *Turkiye Klinikleri*. doi: 10.3906/elk-1907-46.

[47] N. Dankolo *et al.*, "Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words," *Int J Comput Appl*, vol. 185, no. 33, pp. 975–8887, Sep. 2023.

[48] A. W. Andi, C. Slamet, D. S. Maylawati, J. Jumadi, A. R. Atmadja, and M. A. Ramdhani, "Sentiment Analysis of State Capital Relocation of Indonesia using Convolutional Neural Network," in *2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED)*, 2022, pp. 1–6. doi: 10.1109/ICCED56140.2022.10010503.

[49] J. Muliawan and E. Dazki, "SENTIMENT ANALYSIS OF INDONESIA'S CAPITAL CITY RELOCATION USING THREE ALGORITHMS: NAÏVE BAYES, KNN, AND RANDOM FOREST," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 5, pp. 1227–1236, 2023, doi: 10.52436/1.jutif.2023.4.5.347.

[50] K. S. S. Varma, P. Sathineni, and R. Mamidi, "Sentiment Analysis in Code-Mixed Telugu-English Text with Unsupervised Data Normalization," in *International Conference Recent Advances in Natural Language Processing, RANLP*, Incoma Ltd, 2021, pp. 753–760. doi: 10.26615/978-954-452-072-4_086.

[51] S. K. Johal and R. Mohana, "Effectiveness of Normalization Over Processing of Textual Data Using Hybrid Approach Sentiment Analysis," *Int. J. Grid High Perform. Comput.*, vol. 12, pp. 43–56, 2020, doi: 10.4018/ijghpc.2020070103.

[52] A. S. Safitri, I. Wijayanto, and S. Hadiyoso, "Improving Classification Accuracy With Preprocessing Techniques For Sentiment Analysis," *2024 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 487–490, 2024, doi: 10.1109/ICoDSA62899.2024.10651657.

[53] P. Vrnssvsaileela, N. Naga, M. Rao, and A. Budati, “Iterative Ensemble Learning over High Dimensional Data for Sentiment Analysis,” *Scalable Comput. Pract. Exp.*, vol. 25, pp. 1219–1234, 2024, doi: 10.12694/scpe.v25i2.2650.

[54] T. Trueman, A. K. Jayaraman, and E. Cambria, “An N-gram-Based BERT model for Sentiment Classification Using Movie Reviews,” *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 41–46, 2022, doi: 10.1109/AIDE57180.2022.10060044.

[55] A. Zakaria and M. Siallagan, “Predicting Customer Satisfaction through Sentiment Analysis on Online Review,” *International Journal of Current Science Research and Review*, p., 2023, doi: 10.47191/ijcsrr/v6-i1-56.

[56] F. Resyanto, Y. Sibaroni, and A. Romadhony, “Choosing The Most Optimum Text Preprocessing Method for Sentiment Analysis: Case:iPhone Tweets,” in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985943.

[57] S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, “Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data,” in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 2019, pp. 1–8. doi: 10.1109/KSE.2019.8919368.