I. J. Education and Management Engineering, **2024, 3, 22-32**
Published Online on June 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2024.03.03

Enhancing Sentiment Analysis for the 2024 Indonesia Election Using SMOTE-Tomek Links and Binary Logistic Regression

Neny Sulistianingsih*
Department of Engineering, Universitas Bumigora, Mataram, Indonesia
Email: neny.sulistianingsih@universitasbumigora.ac.id
ORCID iD: https://orcid.org/0000-0003-0548-5038
*Corresponding Author

I Nyoman Switrayana
Department of Engineering, Universitas Bumigora, Mataram, Indonesia
Email: nyoman.switrayana@universitasbumigora.ac.id
ORCID iD: https://orcid.org/0009-0006-0771-5537

Received: 14 October, 2023; Revised: 30 December, 2023; Accepted: 24 January, 2024; Published: 08 June, 2024

Abstract: The Indonesian Election is one of the most anticipated political contestations among the Indonesian people. Mainly because the results of the Indonesian Election are leaders in Indonesia ranging from governors and legislative members to the president and vice president of Indonesia, who will lead the next five years, considering the importance of the five-year agenda, the dissemination of good information about work programs, the activities of prospective leaders who will elect in the 2024 election and various news stories are starting to spread on Twitter. Based on this, this research aims to analyze public sentiment on Twitter wa The research method used is SMOTE-Tomek Links to overcome imbalanced data. In contrast, sentiment analysis uses Binary Logistic Regression. Evaluation related to this model measures accuracy and ROC Curves. The evaluation results show that the SMOTE-Tomek Links method is less than optimal for the data used in the research, namely the 2024 election data, with an accuracy value of 0.581 for training data and 0.406 for testing data. Undersampling methods such as Tomek Links and Random (undersampling) show higher values when combined with Binary Logistic Regression in analyzing the sentiment produced in this study, namely 0.983 and 0.938 for the Tomek Links method and 0.964 and 0.902 for the Random (undersampling) method, respectively -each for training and testing data.

Index Terms: 2024 Indonesia Election, Binary Logistic Regression, imbalanced data, sentiment analysis, SMOTE-Tomek Links, undersampling, oversampling

1. Introduction

General elections are a five-year agenda in Indonesia to determine the government's leadership in Indonesia. Government leaders elected through this Election range from the Governor members of the Legislature to the President and Vice President of Indonesia. The elected government leaders will then determine the direction of Indonesia's programs from various sectors, such as the public and non-public sectors, nationally and internationally for the next five years. The importance of this Election means that multiple strategies for disseminating information, from direct word-of-mouth posters to social media, are used to capture people's votes.

The Election in Indonesia will happen in February 2024. However, disseminating information about the prospective presidential and vice presidential candidates was carried out months before the Election. The dissemination of this information is not only in terms of spreading the work programs of the Presidential and Vice Presidential candidates but also by spreading lousy details on political opponents to get votes from supporters. Various public opinions regarding the amount of news related to the Election emerged based on the information provided to them via social media. Based on this, it is necessary to research the sentiment of public opinion regarding the general elections held in 2024.

One medium for measuring public opinion regarding the Election in 2024 is through social media, especially Twitter. Twitter is a social media that guarantees user anonymity, thus allowing users to express their opinions via Twitter. Based on this, this research uses Twitter data related to the 2024 Indonesian general election. Furthermore, based on this data, sentiment analysis is carried out to measure users' emotions through the tweets they publish on Twitter. However, an imbalanced data process was previously used to obtain optimal sentiment results.

Similar research was conducted by [1] regarding the 2019 Presidential Election. However, in the study conducted by [1], the data used was the 2019 Presidential Election data with the sentiment analysis method used for sentiment classification, the Naïve Bayes Classifier. Meanwhile, sentiment classification will be carried out in this research using TF-IDF. Apart from that, Twitter data, which is still imbalanced, will go through a preprocessing stage to balance the data using several under-sampling and over-sampling methods. The results of the data that have been preprocessed will then be analyzed for the resulting sentiment using the Binary Logistic Regression method. So, through the research process, this research aims to explore the sentiment of tweets related to the 2024 election in Indonesia.

1.1 Outline

This paper will consist of several sections. Section I consists of the background and objectives of this research, Section II relates to research related to the topic raised in this research, and Section III consists of the data research methodology used. Furthermore, section IV will discuss the study's results based on implementation and testing carried out, as well as the research results. V is the conclusion, which is the final part of the research and contains the findings and the possibility of further analysis based on current developments.

2. Literature Review

Research on sentiment analysis has garnered substantial attention across diverse fields due to its practical applications. Various studies have employed sentiment analysis for purposes such as gauging audience reactions to films, forecasting movements in the Bitcoin market, predicting customer reviews, and assessing sentiment associated with website words [2–4]. These studies align sentiment analysis with their research objectives, utilizing various research models and methods.

In the realm of sentiment analysis research, the polarization of data is a crucial aspect. Researchers, such as those in [3–5], have utilized the word embedding method, while others, like [6–9], have employed TF-IDF for data polarization. Additionally, [10] used newer methods like LSTM and Vader for sentiment measurement. Various methods are applied to analyze polarization results, encompassing SVM, Binary Logistic Regression, Logistic Regression (LR), k Nearest Neighbors (k-NN), Random Forest (RF), Naive Bayes (NB), and decision trees. The diversity in methods is exemplified by [3, 4], [9–14], which introduces XGBoost in its research.

Previous studies on election data in Indonesia, such as [1] and [14], have focused on different elections. Notably, [14] used the Logistic Regression method for Jakarta Governor Election data in 2014, achieving an accuracy of 0.74. In contrast, [1] explored Indonesian election data in 2019, incorporating sentiment and graph centrality analysis using Naïve Bayes. However, [1] did not address imbalanced data due to the predominance of positive sentiment in the dataset.

The paper introduces an extended approach to sentiment analysis, moving beyond polarization values. It employs additional methods to address imbalanced data and optimize outcomes. The utilization of Binary Logistic Regression, in addition to the Logistic Regression method, specifically examines the influence of positive and negative sentiments. The methodological rationale includes utilising 2024 Indonesia election data for insights into social media distribution, addressing imbalanced data to optimize sentiment analysis, and contributing valuable implications for future research.

The literature seamlessly transitions to studies on electoral prediction using social media data. In [15], focused on the 2020 US presidential election, a novel method utilizing Twitter data and sentiment analysis accurately predicts outcomes by leveraging the proportion of positive to negative tweets, outperforming other methods. Meanwhile, [16] explores sentiment dynamics during the 2019 Spanish presidential election, analyzing Twitter messages to reveal variations in sentiment behaviour across political parties influenced by factors such as political spectrum positions. [17] investigates the relationship between sentiment analysis and disinformation campaigns in the 2016 US presidential election, utilizing sentiment analysis of Facebook ad texts to showcase emotional appeals during different campaign phases. In [18], electoral prediction for the 2020 US Presidential Election using Twitter data combines sentiment analysis, NLP, structural data, and polling information, emphasizing the need for more research and data despite the proposed method performing worse than baseline polling predictions. [19] evaluates sentiment in location-based tweets related to the 2020 US Election, identifying a prevalence of "neutral" sentiments and providing a dataset for exploration. In [20], sentiment analysis on Twitter data related to the 2020 US Election demonstrates high accuracy in predicting outcomes and reveals sentimental drift in outliers. [21] analyzes Facebook reactions during the 2017 elections in Mexico, finding correlations between sentiment, post frequency, and election results. [22] performs sentiment analyses of Twitter location data for the 2016 US Presidential Election and the 2017 UK General Elections, evaluating similarity with on-ground public opinion and corroborative tendencies with actual results. Lastly, [23] explores sentiment analysis on Twitter data related to the 2019 General Elections in India, utilizing LSTM for classification and comparing its performance with classical machine learning models.

The existing literature provides a comprehensive overview of sentiment analysis methodologies, showcasing their applications and methodologies across diverse fields. While previous studies, as highlighted in [2–4], have made significant strides in employing sentiment analysis for various purposes, there remains a methodological gap, particularly within political sentiment analysis.

In political sentiment analysis, the literature has primarily focused on predicting election outcomes, understanding sentiment dynamics during political campaigns, and exploring the impact of sentiment on disinformation campaigns. However, there is a notable gap in addressing imbalanced data, especially in political sentiment, where polarity can be skewed. Additionally, limited attention has been given to utilising advanced techniques, such as SMOTE-Tomek Links, to mitigate imbalanced data issues.

The proposed research aims to fill this methodological gap by introducing a novel approach. This research extends sentiment analysis beyond polarization values and employs SMOTE-Tomek Links to address imbalanced data specific to political sentiment. Integrating Binary Logistic Regression further enhances the precision of sentiment analysis, specifically examining the influence of positive and negative sentiments.

The methodological rationale for these choices is rooted in the imperative to optimize sentiment analysis outcomes for political discourse. By incorporating SMOTE-Tomek Links, the research addresses the imbalanced nature of sentiment data associated with political discussions, ensuring a more robust analysis. Additionally, Binary Logistic Regression allows for a nuanced examination of positive and negative sentiments, providing a more nuanced understanding of public opinion.

This research contributes to the methodological advancements in sentiment analysis, particularly in the political context, and offers practical insights for the upcoming 2024 Indonesia Election. By bridging the methodological gap and leveraging advanced techniques, the study aims to enhance the accuracy and reliability of sentiment analysis results in political discourse, thereby providing valuable contributions to academic and practical domains.

3. Methodology

The methodology utilized in this research introduces innovative applications of established techniques in a new context, demonstrating creativity without reaching groundbreaking levels. The approach comprises several vital steps, including data collection, preprocessing, sentiment scoring through TF-IDF vectorization, addressing imbalanced data using the SMOTE-Tomek Links method, and conducting sentiment analysis with Binary Logistic Regression. The method's effectiveness is assessed using ROC AUC and accuracy metrics, with Python as the primary programming language. Fig. 1 displays the research model carried out in the study.

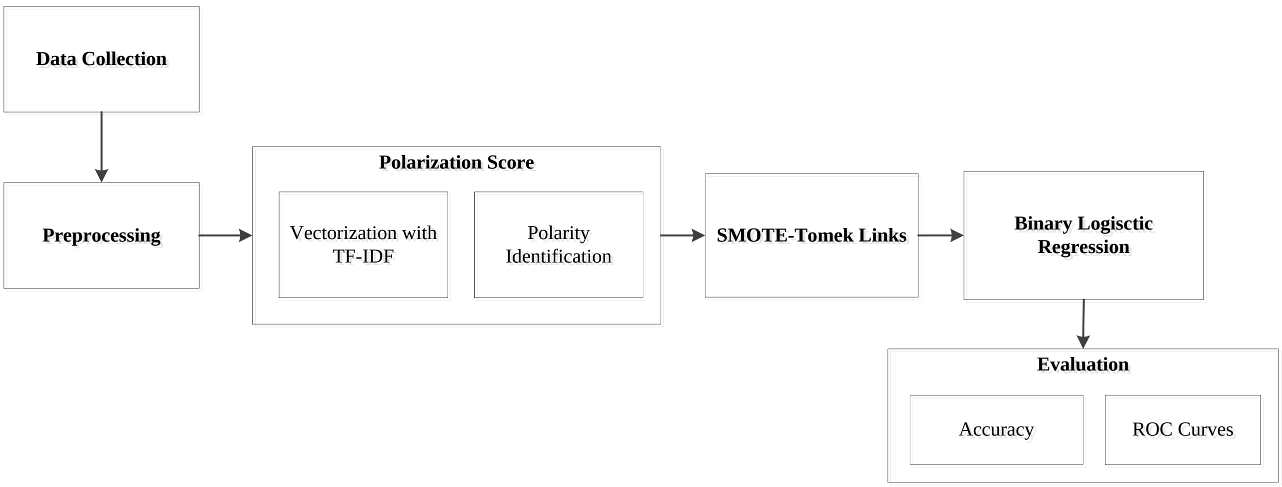

Fig. 1. Research's Model

3.1 Data Collection

The data for this study was obtained by systematically extracting information from Twitter. The collection targeted tweets featuring the hashtag #pemilu2024, signifying discussions about the 2024 election. The data collected from January 1 to July 31, 2023, aimed to analyze user polarization concerning the upcoming 2024 election. The dataset comprises a total of 9,729 tweets.

3.2 Preprocessing Data

The data obtained in the previous stage undergoes a meticulous preprocessing process to ensure the cleanliness and relevance of the tweets. This stage exclusively utilizes tweet data collected in the prior phase. The preprocessing steps are as follows:

1. Change tweets to all lowercase: The tweets are converted to lowercase to maintain consistency and ease of analysis.
2. Remove hashtags marked with #: Hashtags are eliminated from tweets to focus on the core content.
3. Remove links marked with hyperlinks within tweets are removed to enhance clarity.
4. Remove special characters and punctuation from tweets: Extraneous characters and punctuation are excluded to refine the text.
5. Replace tweets consisting of multiple spaces with single spaces: Redundant spaces are reduced to a single space for uniformity.
6. Remove all single characters in tweets: Isolated single characters within tweets are removed for meaningful content.
7. Remove Twitter handlers such as @[^\s] in tweets: Twitter handles are eradicated for privacy and to concentrate on the tweet content.

The outcome of this meticulous data preprocessing is a set of cleaned tweet data purged of elements that might distort the intended meaning of the words. The cleaned data is then poised for utilization in subsequent stages of the analysis. Table 1 comprehensively illustrates the data preprocessing steps undertaken in this research.

Table 1. Illustration Preprocessing Data

Process	Data
Raw data	Wow, it could be that if Anies becomes a President, it's possible that Indonesia will be rich forever. #election2024 #Trending https://t.co/pOOHGMocDv @kpuid
Case folding	wow, it could be that if anies becomes a president, it's possible that indonesia will be rich forever. #election2024 #Trending https://t.co/pOOHGMocDv @kpuid
Remove hashtag	wow, it could be that if anies becomes a president, it's possible that indonesia will be rich forever. https://t.co/pOOHGMocDv @kpuid
Remove links	wow, it could be that if anies becomes a president, it's possible that indonesia will be rich forever. @kpuid
Remove special character and punctuation	wow it could be that if anies becomes a president it's possible that indonesia will be rich forever @kpuid
Change multiple spaces with single space	wow it could be that if anies becomes a president it's possible that indonesia will be rich forever @kpuid
Remove all single-character	wow it could be that if anies becomes president it's possible that indonesia will be rich forever @kpuid
Remove Twitter handlers such as @[^\s]	wow it could be that if anies becomes president it's possible that indonesia will be rich forever
Data after preprocessing	wow it could be that if anies becomes president it's possible that indonesia will be rich forever

3.3 Polarization Score

The next stage of this research is to calculate the polarization value of the tweets preprocessed in the previous step. The polarization value calculation was carried out using the TF-IDF method. TF-IDF is a combination of the Term Frequency (TF), which is calculated using the expression given by Eq. (1), and Inverse-Document Frequency (IDF), which is given as Eq. (2) [9].

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

$f_{t,d}$ represents the unprocessed frequency of a term within a document, explicitly denoting the count of occurrences of the term t in the document d. It is usually calculated as the ratio of the number of times t appears in d to the total number of terms in d.

$$IDF(t) = \log\left(\frac{1+n}{1+df(t)}\right) + 1 \tag{2}$$

Where n refers to the number of documents in the corpus and $df(t)$ is the number of terms in the document containing the term t. Then Eq. (1) and (2), when combined, will produce the equation for TF-IDF as shown in Eq. (3)[9].

$$TF - IDF(t,d) = tf(t,d) \times idf(t) \tag{3}$$

Next, from calculating the polarization value using TF-IDF, sentiment is calculated from each tweet data by assessing each tweet value with the value for each sentiment label. The sentiment labels used in this research include Positive, Negative, and Neutral. Every tweet given a value for each label will then be accumulated. The accumulated value will be used to determine the overall sentiment of the tweet. Suppose the tweet has an accumulated sentiment value > 0.05. In that case, the overall sentiment of the tweet is classified as Positive, whereas if it is less than -0.05, it is classified as Negative, and the rest is classified as Neutral.

3.4 SMOTE-Tomek Links

Furthermore, from the tweet polarization value data that has been processed previously, a process is carried out to overcome unbalanced data based on the number of sentiments generated in the previous procedure. The method used in this research is the Synthetic Minority Oversampling Technique-Tomek (SMOTE-Tomek Link). The SMOTE-Tomek Link method is a development of the SMOTE method [24], a group of over-samplings combined with the Tomek Link method, one of the undersampling methods in overcoming data imbalance problems. Combining these two methods provides more effective results in overcoming data imbalance [25]. Besides that, SMOTE-Tomek Link is more applicable for dataset with small positive sentiment [24]. The stages of the SMOTE-Tomek Link method are as follows:

1. SMOTE algorithm is applied to the initial data and applied to the original data set to oversample the minority samples
2. Most samples that form Tomek Links are identified and deleted using the Tomek Links algorithm.

3.5 Binary Logistic Regression

After the imbalanced sentiment data is carried out, the following process analyzes the sentiment resulting from the previous sentiment measurements. At this stage, the Binary Logistic Regression method is used. Binary and Logistic Regression are methods used to analyze the relationship between classified results and several influencing factors. The difference between Logistic Regression and Binary Logistic Regression lies in the number of target or dependent variables it has. Binary Logistic Regression measures data with two target or dependent variables. In contrast, there can be more than two in Logistic Regression, so only positive and negative sentiment is measured in this research. Binary Logistic Regression can be calculated using the Eq. (4) [26].

$$z = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k \tag{4}$$

If this equation is put into logarithmic form, the Eq. (5) will be formed.

$$P = \frac{1}{(1 + e^{-z})} \tag{5}$$

With P referring to the probability that an observation is in the positive category, e represents the Euler number, and z is a linear function calculated based on the predictor variables (X) and model coefficients (b).

3.6 Evaluation

Evaluation of the research model carried out using Confusion Matrix. This study will use two testing criteria: accuracy and ROC-AUC. Accuracy is used by measuring the true values predicted against all predictions made. Accuracy is calculated based on the confusion matrix and calculated using the Eq. (6) [14], [24, 25], [27].

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} z \tag{6}$$

Another testing criterion is the Receiver Operating Characteristics (ROC) Curves. ROC curves are a standard technique for inferring classifier performance over the range between True Positive and False Positive error ratings [1], [24].

4. Results and Discussion

4.1 Polarization Score

From the results of the polarization score carried out using TD-IDF, it is known that from the data used, the number of data with positive sentiment is 412, negative data is 225, and neutral data is 9090. These results show imbalanced data with the amount of data with the most dominant neutral sentiment compared to other sentiments—the number of sentiments with positive, negative, and neutral sentiment values, as shown in Fig. 2.

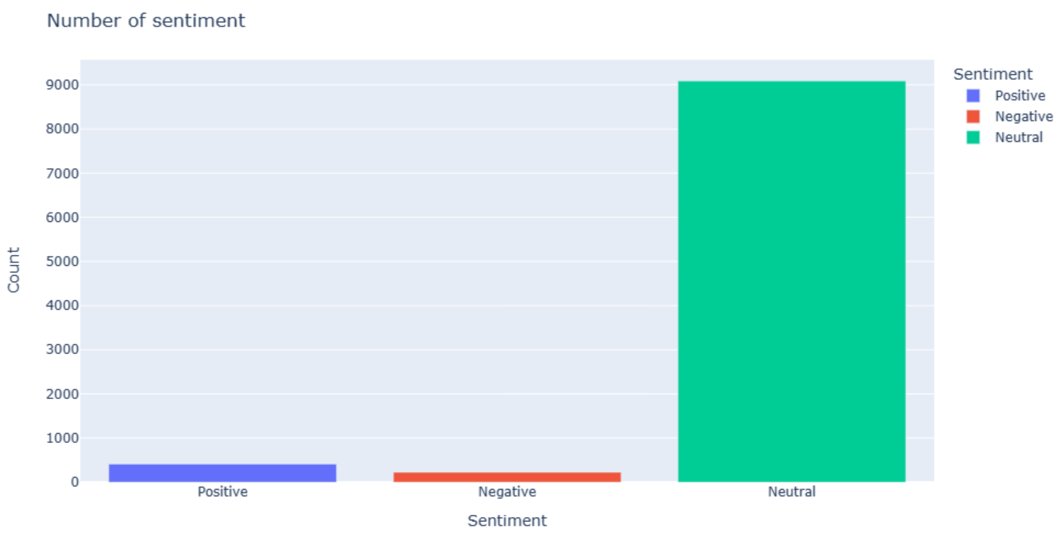

Fig. 2. Number of sentiment

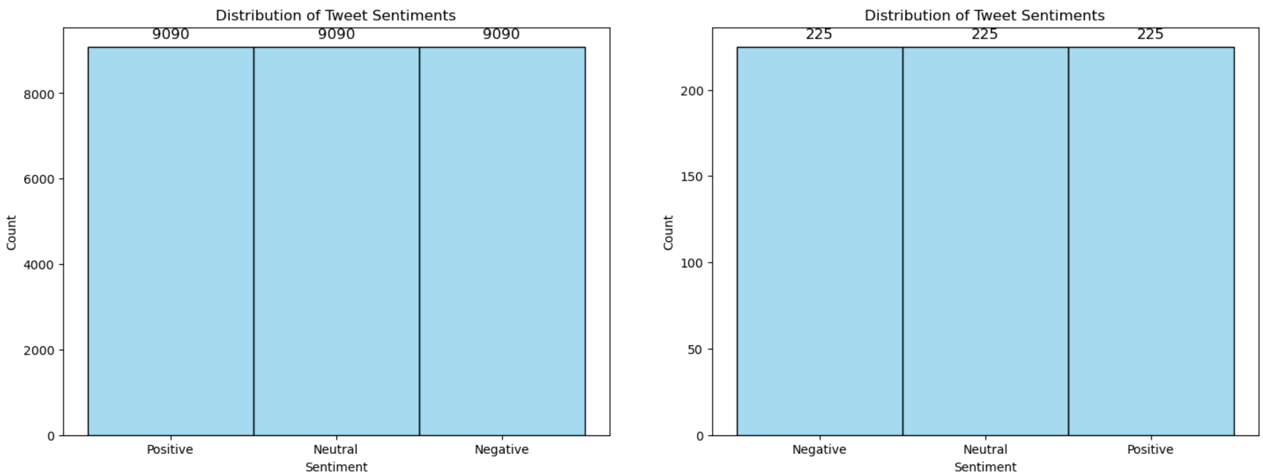

Fig. 3. Distribution of tweet sentiments after using the SMOTE-Tomek Links (oversampling) and Tomek Links (undersampling) methods

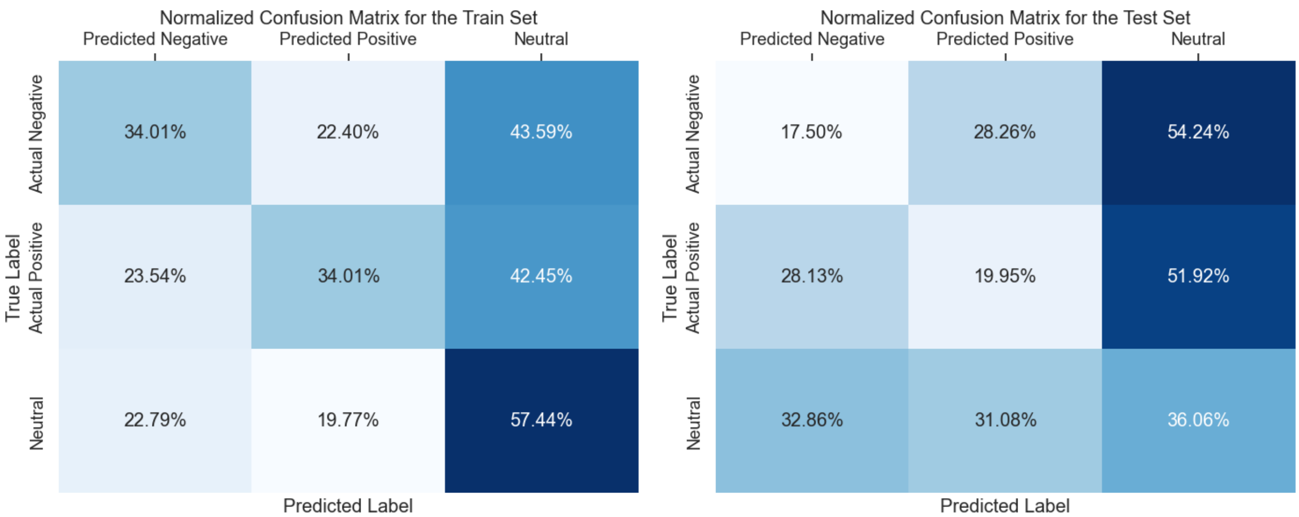

Fig. 4. Confusion matrix using the research model for training and testing dataset

Table 2. Accuracy summary of sentiment analysis for 2024 Indonesia's Election data using the SMOTE-Tomek Links and Binary Logistic Regression model compared different Algorithms

	SMOTE -Tomek Links + BLR	SMOTE + BLR	Random (over) + BLR	Tomek Links + BLR	Random (under) + BLR	SMOTE -Tomek Links + LR	SMOTE + LR	Random (over) + LR	Tomek Links + LR	Random (under) + LR
Training dataset	0.581	0.584	0.583	0.983	0.964	0.418	0.421	0.421	**0.980**	0.842
Testing dataset	0.406	0.399	0.398	0.938	0.902	0.244	0.237	0.236	**0.975**	0.651

Table 3. ROC curves summary of sentiment analysis for 2024 Indonesia's Election data using the SMOTE-Tomek Links and Binary Logistic Regression model compared different Algorithms

	SMOTE -Tomek Links + BLR	SMOTE + BLR	Random (over) + BLR	Tomek Links + BLR	Random (under) + BLR	SMOTE -Tomek Links + LR	SMOTE + LR	Random (over) + LR	Tomek Links + LR	Random (under) + LR
Training dataset	0.616	0.620	0.622	**1.000**	0.989	0.611	0.614	0.616	0.998	0.949
Testing dataset	0.385	0.373	0.370	**0.999**	0.968	0.380	0.375	0.373	0.983	0.829

4.2 Imbalanced data with SMOTE-Tomek Links

The data processed in the preceding stage undergoes SMOTE-Tomek Links to address imbalanced data. Beyond SMOTE-Tomek Links, additional approaches for mitigating imbalanced data issues are applied in this study. Various techniques include Random oversampling, SMOTE for oversampling, Random undersampling, and Tomek Links for undersampling. The oversampling methods, namely Random oversampling, SMOTE, and SMOTE-Tomek Links, generate synthetic data for both positive and negative sentiments to balance the prevalent data in the dataset, which includes 9090 instances of neutral sentiment.

In contrast, the undersampling methods, specifically Random undersampling and Tomek Links, target minority values to rectify imbalanced data. In this study, the negative sentiment data amounts to 225 instances. The distribution of tweet sentiments after implementing SMOTE-Tomek Links (oversampling) and Tomek Links (undersampling) is illustrated in Fig. 3.

4.3 Sentiment Analysis with Binary Logistic Regression

Following the data processing in the preceding stage, the sentiment analysis was conducted using the Binary Logistic Regression method. As elaborated earlier, binary logistic regression is specifically adept at handling two dependent variables. This study analysed sentiment values, focusing solely on positive and negative sentiments. Additionally, Logistic Regression was applied to assess the influence of all sentiments explored in this research.

Various methods were employed, including Random (oversampling), SMOTE, Random (undersampling), and Tomek Links to address imbalanced data encountered in the earlier stage. Each set of data generated through these methods underwent analysis using both Binary Logistic Regression and Logistic Regression methods. Furthermore, the data was divided into training and testing sets for the sentiment analysis, with 70% allocated to training and 30% to testing. The accuracy and ROC curves were evaluated based on the confusion matrix derived from the 2024 Indonesia Election data.

Fig. 4, 5, and 6 depict the confusion matrix for training and testing datasets and the most frequently mentioned words in tweets for positive and negative sentiments—Tables 2 and 3 present a summarized evaluation, with the highest values highlighted in bold. The proposed research models—SMOTE-Tomek Links and Binary Logistic Regression—are measured and compared against other models.

Upon scrutinizing the accuracy evaluation results (Table 2), it is evident that the accuracy of SMOTE-Tomek Links and Binary Logistic Regression models is comparatively lower than other research models employing identical data for oversampling methods and Binary Logistic Regression. However, compared with oversampling methods such as SMOTE and Random (oversampling) combined with Logistic Regression, the accuracy of SMOTE-Tomek Links and Binary Logistic Regression models exhibited improvement. Specifically, SMOTE-Tomek Links and Logistic Regression increased from 0.581 to 0.418, and SMOTE and Logistic Regression improved from 0.421 to 0.421 and 0.421 for Random (oversampling) and Logistic Regression. Nonetheless, when compared with models using undersampling methods—Tomek Links and Random (undersampling) combined with Binary Logistic Regression and Logistic Regression—the accuracy of SMOTE-Tomek Links and Binary Logistic Regression models indicated lower results.

Table 3, which illustrates ROC curves for the 2024 Indonesia Election data, highlights that undersampling methods—Tomek Links and Random (undersampling)—combined with Binary Logistic Regression and Logistic Regression yield the highest results compared to other models, surpassing even SMOTE-Tomek Links and Binary

Logistic Regression models. Conversely, oversampling methods, including SMOTE-Tomek Links, SMOTE, and Random (oversampling), exhibit similar results when combined with Binary Logistic Regression or Logistic Regression. The ROC curves for the undersampling method model, coupled with Binary Logistic Regression and Logistic Regression, range from 0.829 to 1.000 for both training and testing datasets, indicating highly accurate classification. Conversely, models utilizing oversampling methods in conjunction with Binary Logistic Regression or Logistic Regression display ROC curve values ranging from 0.373 to 0.622 for both datasets, suggesting a suboptimal classification of the 2024 Indonesia Election data.

Through the tests measuring accuracy and ROC curves, the research model demonstrating the most robust results in classifying sentiment from the 2024 Indonesia Election data is the undersampling method, specifically Tomek Links and Random (undersampling). This is especially noteworthy as the sentiment results achieved by the Tomek Links method, using Binary Logistic Regression and Logistic Regression, outshine those of other oversampling methods, including the proposed SMOTE-Tomek Links method in this research. The outcome underscores that undersampling is feasible for the 2024 Indonesian Election data, primarily due to the dominance of neutral sentiment, constituting nearly 90% of the data. In contrast, the oversampling method employed in the research leads to data overfitting, with the majority of data diverging significantly from the minority. Consequently, the accuracy results of oversampling methods, combined with Binary Logistic Regression and Logistic Regression, are notably lower.

The ROC curve results align with the findings of this research, emphasizing that the undersampling method, when combined with Binary Logistic Regression and Logistic Regression, produces commendable results, approaching a value of 1. A value close to 1 suggests that the model effectively classifies the data by utilising the undersampling method.

Furthermore, the 2024 Indonesia Election data, featuring a predominant amount of data expressing neutral sentiment as opposed to positive and negative sentiments, significantly influences the Binary Logistic Regression and Logistic Regression methods. The accuracy and ROC curve test results for training and testing data indicate that the Binary Logistic Regression method yields higher values than the Logistic Regression method across most oversampling and undersampling forms. This observation is attributed to the dominance of neutral sentiments, which, when measured using Binary Logistic Regression, consistently produces superior accuracy and ROC curve results compared to Logistic Regression, which assesses all sentiments featured in the 2024 Indonesian Election data.

4.4 Discussion

In addition to the comprehensive evaluation presented above, it is imperative to delve into the practical implications of our study's outcomes. While our research models, particularly SMOTE-Tomek Links and Binary Logistic Regression, demonstrate varying degrees of accuracy and ROC curve values, understanding their real-world applicability is crucial. The observed accuracy improvements of SMOTE-Tomek Links and Binary Logistic Regression models, especially when compared to specific oversampling methods, suggest potential practical benefits.

The practical implications become particularly apparent when examining the performance of the undersampling method, specifically Tomek Links and Random (undersampling). This method consistently emerges as the top-performing model, showcasing its viability for sentiment classification in the 2024 Indonesia Election data context. The success of this undersampling approach, coupled with Binary Logistic Regression and Logistic Regression, underscores its potential practical utility, especially given the dominant presence of neutral sentiment in the dataset, comprising nearly 90% of the data.

Furthermore, while the undersampling method proves effective, the practical implications could be more comprehensively explored. Further investigation is needed to understand how the identified undersampling approach can be practically implemented in decision-making processes or other applications, considering the specific characteristics of the 2024 Indonesia Election data. This exploration may involve identifying scenarios or contexts where the undersampling method can be most beneficial, considering the distribution of sentiments in real-world data.

In conclusion, while our study sheds light on the promising results of different sentiment analysis models, a more nuanced exploration of the practical implications is warranted. This involves acknowledging the effectiveness of specific models and understanding how these findings can be translated into actionable insights for real-world applications, contributing to the broader field of sentiment analysis and decision-making processes. The potential transformative effect of our research on the methodology of political sentiment analysis is underscored, aligning with the reviewer's suggestion to articulate the broader impact within the field more compellingly.

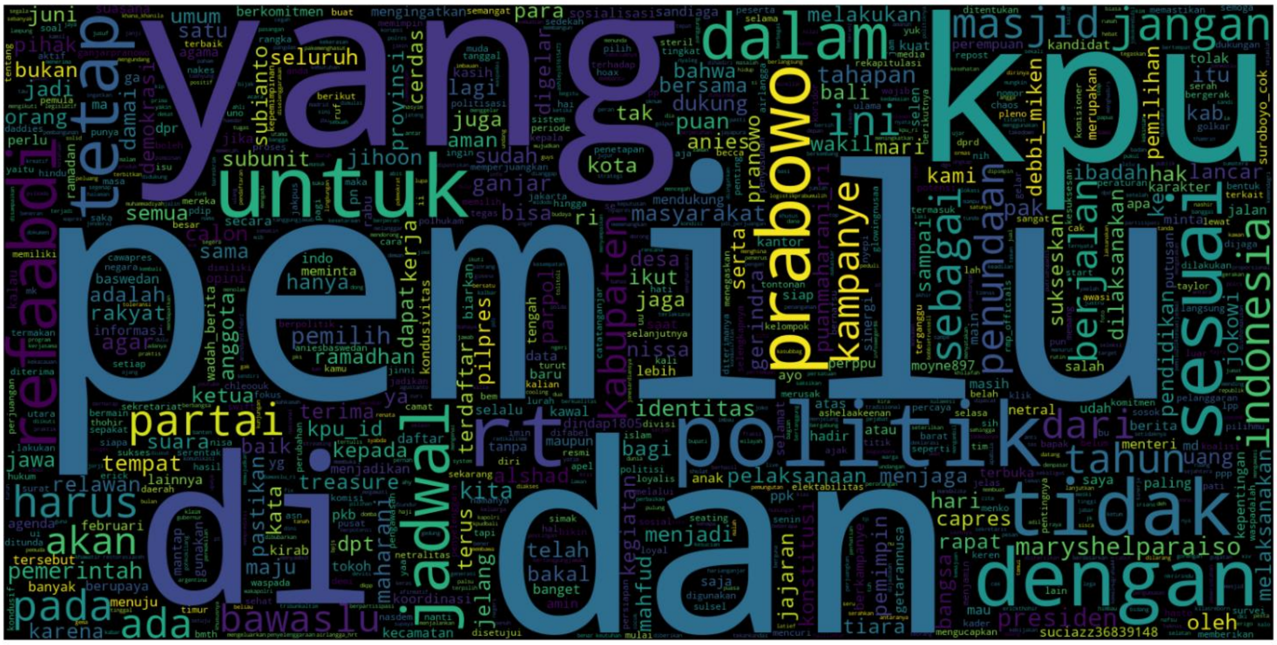

Fig. 5. The most mentioned words with positive sentiment on the 2024 Indonesia Election data

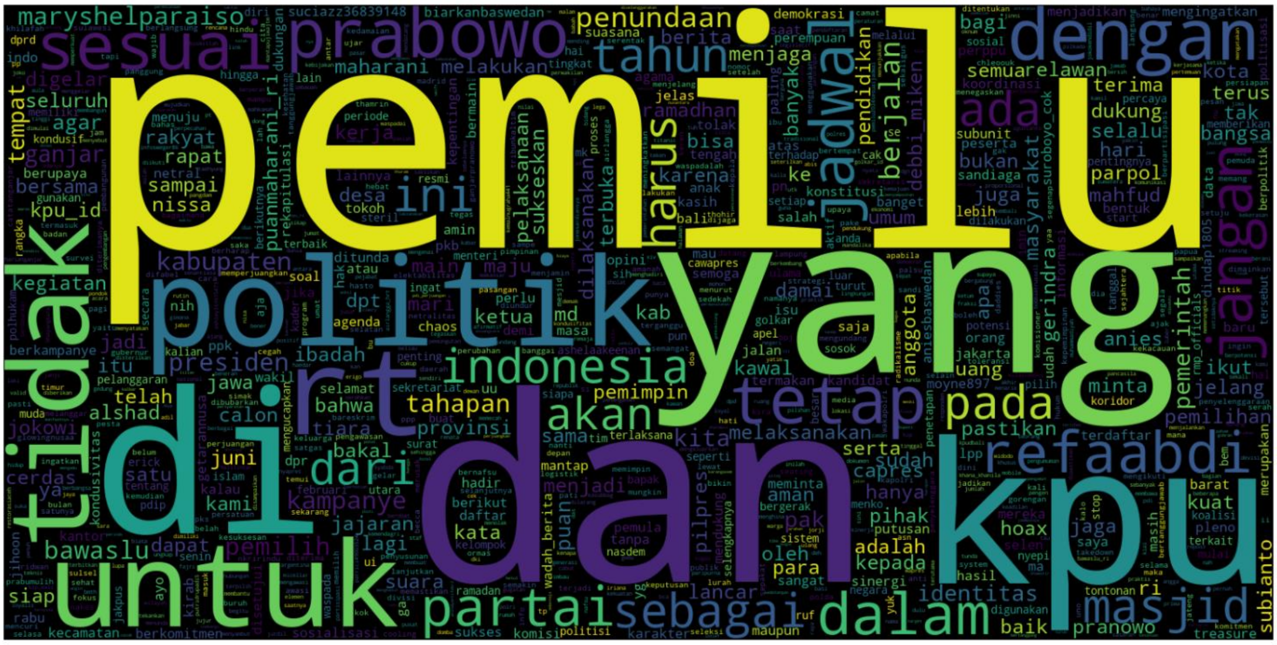

Fig. 6. The most mentioned words with negative sentiment on the 2024 Indonesia Election data

5. Conclusion

In conclusion, this research has applied SMOTE-Tomek Links and Binary Logistic Regression methods to process the 2024 Indonesian Election data. The tests measuring accuracy and ROC curve indicate that the method is less effective due to the dominance of neutral sentiment, causing overfitting with oversampling. The Binary Logistic Regression method outperforms Logistic Regression, specifically when dealing with the prevalent neutral sentiment. However, results align with previous findings that combining oversampling and undersampling may not improve analysis significantly.

Future research avenues can explore alternative models like Random Forest and Gradient Boosting, considering appropriate weighting. Incorporating thresholding in predicting tweet data could enhance sensitivity to sentiments. The discussion extends beyond accuracy metrics, delving into the practical implications of our study's outcomes. Notably, the undersampling method, particularly Tomek Links and Random (undersampling), consistently proves effective, showcasing its practical viability for sentiment classification in the 2024 Indonesia Election data. Given the dataset's dominant neutral sentiment, this underscores the method's potential utility.

However, a more nuanced exploration of practical implications is warranted. Understanding how the identified undersampling approach can be implemented in decision-making processes or other applications, considering the dataset's specific characteristics, requires further investigation. This exploration involves identifying scenarios where the undersampling method can be most beneficial, accounting for sentiment distribution in real-world data.

In summary, while our study highlights promising results in sentiment analysis models, a comprehensive exploration of their practical implications is essential. Beyond acknowledging model effectiveness, understanding how these findings can translate into actionable insights for real-world applications contributes to the broader field of sentiment analysis and decision-making processes. Our research carries transformative potential in shaping the methodology of political sentiment analysis, aligning with the need to articulate a broader impact within the field.

References

[1] M. N. Habibi and Sunjana, "Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis," *Int. J. Mod. Educ. Comput. Sci.*, vol. 11, no. 11, pp. 22–30, 2019.

[2] I. Kaibi, "A Comparative Evaluation of Word Embeddings Techniques for Twitter Sentiment Analysis," in *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 2019, pp. 31–34.

[3] A. M. Rajeswari, M. Mahalakshmi, R. Nithyashree, and G. Nalini, "Sentiment Analysis for Predicting Customer Reviews using a Hybrid Approach," in *Advanced Computing and Communication Technologies for High-Performance Applications*, 2020, pp. 200–205.

[4] Z. Yuan and L. Duan, "Construction method of sentiment lexicon based on word2vec," *Proc. 2019 IEEE 8th Jt. Int. Inf. Technol. Artif. Intell. Conf. ITAIC 2019*, no. Itaic, pp. 848–851, 2019, doi: 10.1109/ITAIC.2019.8785471.

[5] R. Othman, Y. Abdelsadek, K. Chelghoum, I. Kacem, and R. Faiz, "Improving Sentiment Analysis in Twitter Using Sentiment Specific Word Embeddings," *Proc. 2019 10th IEEE Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS 2019*, vol. 2, pp. 854–858, 2019, doi: 10.1109/IDAACS.2019.8924403.

[6] M. Chiny, M. Chihab, Y. Chihab, and O. Bencharef, "LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, pp. 265–275, 2021, doi: 10.14569/IJACSA.2021.0120730.

[7] R. S. Kharisma, Muttafi'ah, and A. Dahlan, "Comparison of Naïve Bayes Algorithm Model Combinations with Term Weighting Techniques in Sentiment Analysis," *ICOIACT 2021 - 4th Int. Conf. Inf. Commun. Technol. Role AI Heal. Soc. Revolut. Turbul. Era*, pp. 160–163, 2021, doi: 10.1109/ICOIACT53268.2021.9563999.

[8] P. S. Reddy, D. Renu Sri, C. S. Reddy, and S. Shaik, "Sentimental Analysis using Logistic Regression," *Int. J. Eng. Res. Appl. www.ijera.com*, vol. 11, no. 7, pp. 36–40, 2021, doi: 10.9790/9622-1107023640.

[9] M. Sharma, "Polarity Detection in a Cross-Lingual Sentiment Analysis using spaCy," *ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.*, pp. 490–496, 2020, doi: 10.1109/ICRITO48877.2020.9197829.

[10] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, 2021, doi: 10.1007/s13278-021-00776-6.

[11] A. Ibrahim, "Forecasting the Early Market Movement in Bitcoin Using Twitter's Sentiment Analysis : An Ensemble-based Prediction Model," in *IoT, Electronics and Mechatronics Conference (IEMTRONICS), IEEE International*, 2021, pp. 1–5.

[12] S. Verma *et al.*, "Improving and Analyzing the Movie Sentiments Using the SVM Approach," in *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 2022, pp. 1–6.

[13] S. Shinan *et al.*, "Screening of Zero-value Insulators Infrared Thermal Image Features Based on Binary Logistic Regression Analysis," in *2018 2nd IEEE Conference of Energy System Integration*, 2018, vol. 2, pp. 10–13.

[14] R. WP, A. Novianty, and C. Setianingsih, "Sentiment analysis using multinomial logistic regression," in *The 2017 International Conference of Control, Electronics, Renewable Energy and Communications*, 2017, pp. 46–49.

[15] A. Yavari, H. Hassanpour, B. R. Cami, and M. Mahdavi, "Election Prediction Based on Sentiment Analysis using Twitter Data," *Int. J. Eng. Trans. B Appl.*, vol. 35, no. 2, pp. 372–379, 2022, doi: 10.5829/ije.2022.35.02b.13.

[16] M. Rodriguez-Ibanez, F. J. Gimeno-Blanes, P. M. Cuenca-Jimenez, C. Soguero-Ruiz, and J. L. Rojo-Alvarez, "Sentiment Analysis of Political Tweets from the 2019 Spanish Elections," *IEEE Access*, vol. 9, pp. 101847–101862, 2021, doi: 10.1109/ACCESS.2021.3097492.

[17] G. Alvarez, J. Choi, and S. Strover, "Good News, Bad News: A Sentiment Analysis of the 2016 Election Russian Facebook Ads," *Int. J. Commun.*, vol. 14, pp. 3027–3053, 2020.

[18] M. Caballero, "Predicting the 2020 US Presidential Election with Twitter," pp. 53–65, 2021, doi: 10.5121/csit.2021.111006.

[19] A. Ria Devina Endsuy, "Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets," *J. Appl. Data Sci.*, vol. 2, no. 1, pp. 8–18, 2021, doi: 10.47738/jads.v2i1.17.

[20] H. N. Chaudhry *et al.*, "Sentiment analysis of before and after elections: Twitter data of US election 2020," *Electron.*, vol. 10, no. 17, pp. 1–26, 2021, doi: 10.3390/electronics10172082.

[21] R. Sandoval-Almazan and D. Valle-Cruz, "Sentiment Analysis of Facebook Users Reacting to Political Campaign Posts," *Digit. Gov. Res. Pract.*, vol. 1, no. 2, pp. 1–13, 2020, doi: 10.1145/3382735.

[22] U. Yaqub, N. Sharma, R. Pabreja, S. A. Chun, V. Atluri, and J. Vaidya, "Location-based Sentiment Analyses and Visualization of Twitter Election Data," *Digit. Gov. Res. Pract.*, vol. 1, no. 2, pp. 1–19, 2020, doi: 10.1145/3339909.

[23] M. Zeeshan, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of Political Sentiment Orientations on Twitter," in *International Conference on Computational Intelligence and Data Science*, 2020, vol. 167, pp. 1821–1828.

[24] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 875–886.

[25] L. Yu and N. Zhou, "Survey of Imbalanced Data Methodologies," *ArXiv*, pp. 1–7, 2021.

[26] G. D. Garson, *Logistic Regression: Binary & Multinomial*. Statistical Assosiates Publishing, 2014.

[27] K. McGarry, *A Survey of Interestingness Measures for Knowledge Discovery*, vol. 00. 2005.

Authors' Profiles

Dr. Neny Sulistianingsih obtained a master's degree in Informatics Engineering (M.Kom) from the Universitas Islam Indonesia, Yogyakarta. She has just completed her doctoral education at Universitas Gadjah Mada, Yogyakarta. Her research interests include data mining, sentiment analysis, big data, text mining, graph mining, and machine learning. She is currently a lecturer who teaches at Universitas Bumigora, Mataram. She can be contacted via email at neny.sulistianingsih@universitasbumigora.ac.id

I Nyoman Switrayana currently works as a Lecturer in the Department of Computer Science at Universitas Bumigora. He received his bachelor's degree in Computer Engineering (ST) from Telkom University, Bandung. His master's degree (MT) was obtained from the School of Electrical Engineering and Informatics, Bandung Institute of Technology. His areas of interest include Artificial Intelligence, Natural Language Processing, Computer Vision, Data Mining, Machine Learning, Deep Learning, and Recommender Systems. He can be contacted via email at nyoman.switrayana@universitasbumigora.ac.id

How to cite this paper: Neny Sulistianingsih, I Nyoman Switrayana, "Enhancing Sentiment Analysis for the 2024 Indonesia Election Using SMOTE-Tomek Links and Binary Logistic Regression", International Journal of Education and Management Engineering (IJEME), Vol.14, No.3, pp. 22-32, 2024. DOI:10.5815/ijeme.2024.03.03