TEKNIKA: JURNAL SAINS DAN TEKNOLOGI

Homepage jurnal: http://jurnal.untirta.ac.id/index.php/ju-tek/

Sentiment analysis on public opinion of electric vehicles usage in Indonesia using support vector machine algorithms

***Naufal Avilandi Poedjimartojo*[a], *Dita Pramesti*[a,1], *Riska Yanu Fa'rifah*[a]**

[a]*Informatic System Department, Universitas Telkom, Jalan Telekomunikasi No. 1, Bandung, 40267, Indonesia*

[1]Corresponding author: ditapramesti@telkomuniversity.ac.id

ARTICLE INFO

Article history:
Submitted 9 September 2023
Received 18 September 2023
Received in revised form 13 October 2023
Accepted 20 October 00
Available online on 1 November 2023

Keywords:
Electric vehicle, sentiment analysis, support vector machine (SVM), oversampling, undersampling

Kata kunci:
Kendaraan listrik, analisis sentimen, support vector machine (SVM), *oversampling, undersampling*

ABSTRACT

Technological developments in the automotive industry have experienced significant progress in recent years. Currently, many electric vehicles are being produced as an environmentally friendly alternative to vehicles. The use of electric vehicles has become an intense topic of conversation in society, giving rise to various responses and opinions on Twitter. This research aims to analyze Indonesian people's sentiment regarding using electric vehicles through data collected from Twitter. Sentiment analysis is carried out using a machine-learning approach. The best method for pattern recognition problems is a Support Vector Machine (SVM) to sort each comment into positive or negative sentiments. Meanwhile, SVM classification performance was measured using the Confusion Matrix method. In this research, the Synthetic Minority Over-Sampling Technique (SMOTE) method and the Random Undersampling (RUS) method were used to overcome data imbalance. After the model creation and performance evaluation process, the best model produced was the baseline Support Vector Machine with a data sharing ratio of 70:30 without applying imbalance handling techniques. This model achieved an accuracy of 94.8%, a precision value of 95.5%, a recall value of 99.1%, and an F-1 Score value of 97.2%.

ABSTRAK

Dalam beberapa tahun terakhir, perkembangan teknologi industri otomotif telah mengalami kemajuan yang signifikan. Saat ini, banyak diproduksi kendaraan berbahan bakar listrik sebagai alternatif kendaraan ramah lingkungan. Penggunaan kendaraan listrik telah menjadi perbincangan intens dalam masyarakat, menimbulkan berbagai macam tanggapan dan pendapat, salah satunya di media sosial Twitter. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat Indonesia mengenai penggunaan kendaraan listrik, melalui data yang dikumpulkan dari media sosial Twitter. Analisis sentimen dilakukan dengan menggunakan pendekatan *machine learning* dan metode terunggul yang dapat diterapkan dalam permasalahan pengenalan pola yaitu *Support Vector Machine* (SVM) untuk memilah setiap komentar menjadi sentimen positif atau negatif. Sedangkan kinerja klasifikasi SVM diukur menggunakan metode *confusion matrix*. Dalam penelitian ini digunakan metode *Synthetic Minority Over-Sampling Technique* (SMOTE) dan metode *Random Undersampling* (RUS) untuk mengatasi ketidakseimbangan data. Setelah proses pembuatan model dan evaluasi performa dilakukan, model terbaik yang dihasilkan adalah *baseline Support Vector Machine* dengan rasio pembagian data 70:30 tanpa penerapan teknik *imbalance handling*. Model ini mencapai hasil akurasi sebesar 94,8%, nilai *precision* sebesar 95,5%, nilai *recall* sebesar 99,1%, dan nilai F-1 *Score* sebesar 97,2%.

Avalaible http://dx.doi.org/10.36055/tjst.v19i2.21967

1. Introduction

Climate change can no longer be avoided and has become a serious issue being discussed by many countries today. Damage that occurs in all ecosystems, natural disasters, extreme weather changes, and depletion of the ozone layer are caused by climate change. One option to reduce the impact caused by climate change is by reducing the use of fossil energy and reducing the amount of carbon dioxide concentration in the air and the level of greenhouse gas emissions.

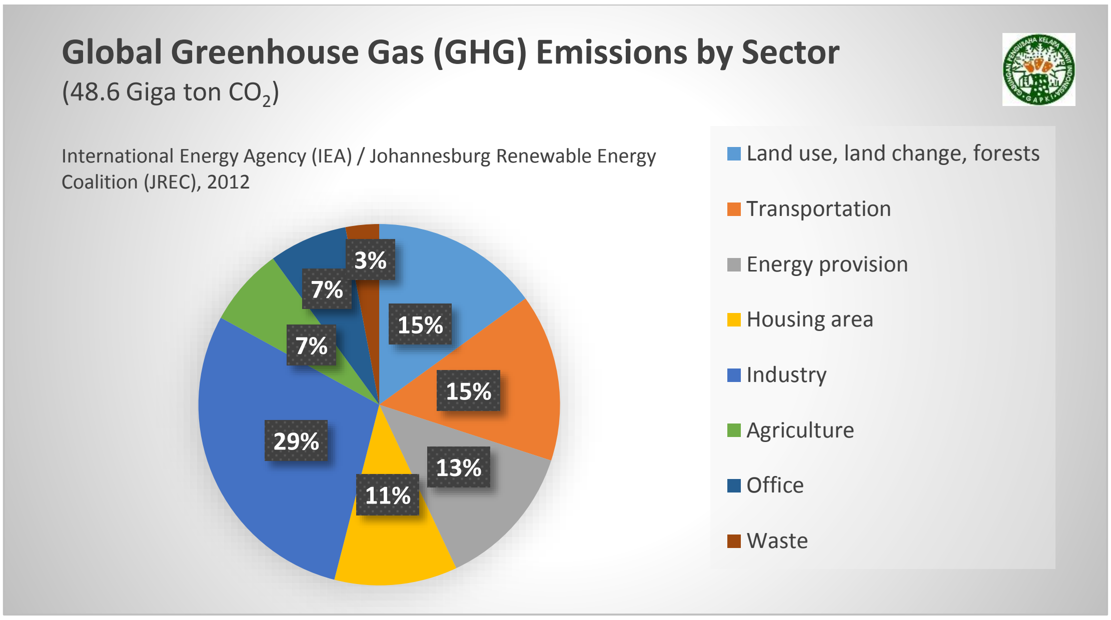


Figure 1. Graph of global greenhouse gas emissions by sector (Gapki.id, 2021).

Figure 1 presents a graph regarding the percentage of global greenhouse gas emissions from various sectors in 2012. The transportation sector contributes 15% of greenhouse gas emissions along with the land use, land change, and forest sectors and is below the industrial sector. Therefore, countries in the world must work together to immediately take firm decisions in dealing with climate change by supporting all activities that have a good impact on the sustainability of various lives.

This issue should also be applied in Indonesia because sooner or later, electric vehicles must become the people's choice as a means of transportation and leave vehicles that still use fossil energy behind. Electric vehicles have several advantages when compared to fossil fuel vehicles. First, electric vehicles have a low engine noise level because they do not use an internal combustion engine like fossil fuel vehicles. Second, electric vehicles have no exhaust because there are no exhaust emissions. Moreover, electric vehicles have high energy efficiency because electricity can be used directly to drive vehicles without wasting much energy in the conversion process [1]. The form of government support has also been stated in Presidential Regulation Number 55 of 2019 concerning accelerating the battery-based road transportation program. This presidential regulation also describes incentives. The use of information technology in Indonesia is increasing. The Indonesian people use information technology to dig deeper into information about electric vehicles on various social media. Electric vehicles are now a hot topic on Twitter and social media. The presence of electric vehicles raises many public responses in the form of opinions that are both pros and cons.

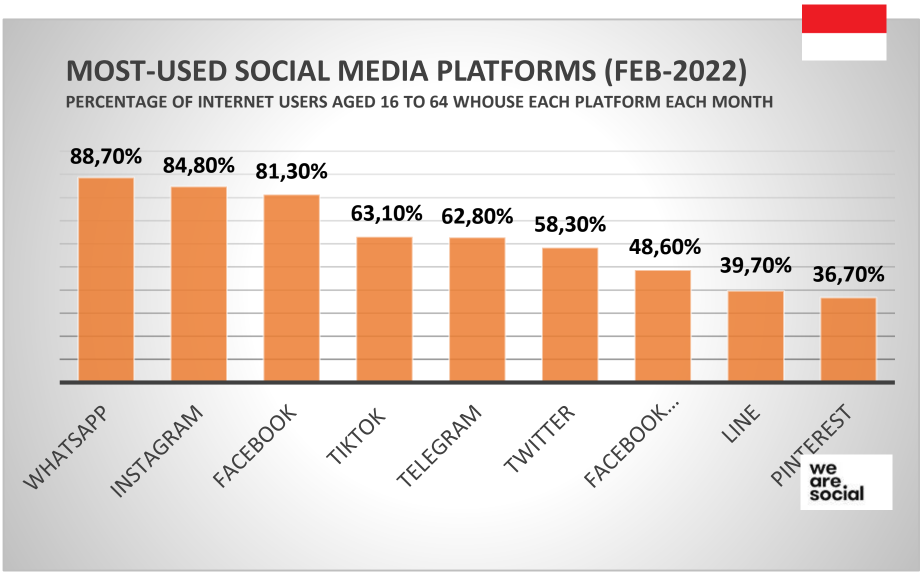


Figure 2. Graphics of most used social media platforms in indonesia in 2022 (we are social, 2022).

Figure 2 depicts that Twitter occupies the sixth position as Indonesia's most widely used social media. Its users reach 58.3% of all active social media users. With so many active users on Twitter and social media, there is also much data regarding public comments regarding the use of electric vehicles. Researchers use the opportunity from this large amount of data to collect and process data, or what is usually called data mining, which aims to extract information to accelerate the transition to electric-based vehicles. The data that has been collected must be used as an explanatory, confirmatory, and exploratory tool. This data mining can also be done through various methods, including association, classification, regression, and clustering. Apart from data mining, researchers also conducted sentiment analysis on public comments regarding the use of electric vehicles to classify public perceptions of whether these comments were positive or negative. Sentiment analysis has several important purposes and benefits in various business, research, and decision-making contexts [2]. One of the goals is to make decisions based on data. Sentiment analysis provides evidence-based data that can be used in decision-making. This issue can help companies plan marketing strategies, promotional campaigns, and product development.

Sentiment analysis is a method for identifying and analyzing opinions or opinions related to a subject, entity, or product in a dataset [3]. Sentiment Analysis, often called opinion mining, entails an automated process of comprehending, extracting, and processing textual data to extract sentiment-related information from opinion sentences [4]. Typically, sentiment analysis involves classification methods to identify and classify opinions or texts into various sentiment categories, such as positive, negative, or neutral. Classification methods help automatically process text and determine how opinions or sentiments are generally expressed. The sentiment analysis in this research uses a machine learning approach with a classification method, namely Support Vector Machine (SVM). Support Vector Machine has a method principle that finds the best separator between different sentiment classes. The Support Vector Machine method contributes to this research because it can be used in all computerized fields. The steps taken were to collect public opinions on Twitter social media regarding electric vehicles, and then these opinions were classified into positive and negative. The results of the information collection process using the Support Vector Machine method are presented in the form of data visualization, which can later be used as recommendation material for the Indonesian government and society as a strategy to support reducing greenhouse gas levels that cause climate change.

The large amount of data available on the internet has triggered rapid growth in efforts to develop information extraction from online databases, one of which is text mining [5]. Text mining, also known as Text Data Mining (TDM) or Knowledge Discovery in Text (KDT), is specifically designed to extract information from unstructured text documents [6]. The central concept behind text mining is mining text data from records to identify words that reflect the document's content [7]. The goal is to analyze the connections between various documents [8]. Text mining aims to overcome the problem of abundant information by applying techniques originating from related fields [9]. Figure 3 below shows the steps of text mining. Figure 3 depicts the text mining processes of tokenizing, filtering, stemming, tagging, and analyzing. It can be considered an extension of data mining or Knowledge Discovery in Databases (KDD), focusing on finding interesting patterns in large databases [10]. The most common stages of text mining are as follows [11].

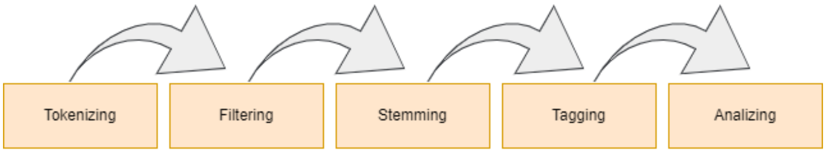


Figure 3. Text mining stages.

Sentiment analysis in Indonesian is a technique for identifying how sentiments are expressed in text and categorizing these sentiments as positive or negative [12]. This issue is also confirmed by Cvijikj & Michahelles (2011), who use sentiment analysis to understand comments made by internet users and see how they receive products or brands. Sentiment analysis is a process for determining opinions, emotions, and attitudes reflected in texts, usually classified into negative and positive views [13]. From the three quotations, sentiment analysis is the process of categorizing sentiments into positive or negative based on the opinions expressed in the text [14]. Internet users often voice their feelings in text form, be it positive, neutral, or negative, which can be expressed in complex ways [15]. The sentiment analysis process involves reviewing collected data, identifying the sentiment, identifying the feature selection, classifying sentiments, and calculating the polarity of the sentiment [16].

A machine learning technique, Support Vector Machine (SVM), examines data and identifies classification and regression analysis patterns [17]. The working principle of SVM is to find the optimal hyperplane with the maximum margin to separate two different classes [18]. In the case of linearly separable classifications, SVM works on this basic principle. However, SVM has developed to work on non-linear problems by adopting the kernel concept in high-dimensional workspaces [19]. The formula used in obtaining the hyperplane is formulated in Equation (1).

$$\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b = 0 \tag{1}$$

where,

w : a vector that has a weight value,

x : the vector that contains the values of the attributes,

b : bias value.

The Synthetic Minority Over-Sampling Technique (SMOTE) is a method in the field of data processing that is used to overcome the problem of class imbalance in classification problems [20]. Class imbalance occurs when the number of samples in one class is much less than in other classes in the dataset [21]. The SMOTE technique takes a sample from the minority label or class and synthesizes the data along a line connecting with one of the nearest neighbors of the minority label [22]. This process successfully balances the data in the training dataset so there are no imbalances during the model-building process [23].

Random Undersampling is a method of reducing the dataset size, which is carried out by randomly deleting several samples from the majority class in the dataset so that the number of samples in the majority class becomes closer to the number of samples in the minority class [24]. The main goal of Random Undersampling is to create a more balanced distribution of classes in the dataset so that the classification model does not become too likely to predict the majority class [25]. However, by reducing the dataset's size, RUS can potentially remove data that contains valuable information and essential patterns in the dataset that should help the classification process and can ultimately reduce the performance of the classification algorithm [26].

Performance evaluation is a metric employed to gauge the precision of algorithm implementation [27]. Several evaluation criteria include accuracy, standard deviation, f1-score, recall, precision, and specificity [28]. In this research, evaluation was carried out by calculating accuracy, precision, recall, and f1-score. The confusion matrix obtains accuracy, precision, recall, and f1-score values [29]. The confusion matrix is a tool that analyses the extent to which a classification model can recognize various data tuples correctly [30].

2. Method

This study's problem-solving systematics consisted of initiation, data processing, and results and conclusions. Knowledge Discovery in Database (KDD) is the problem-solving systematic used in this research. The following is a systematic problem-solving diagram used, presented in Figure 4.

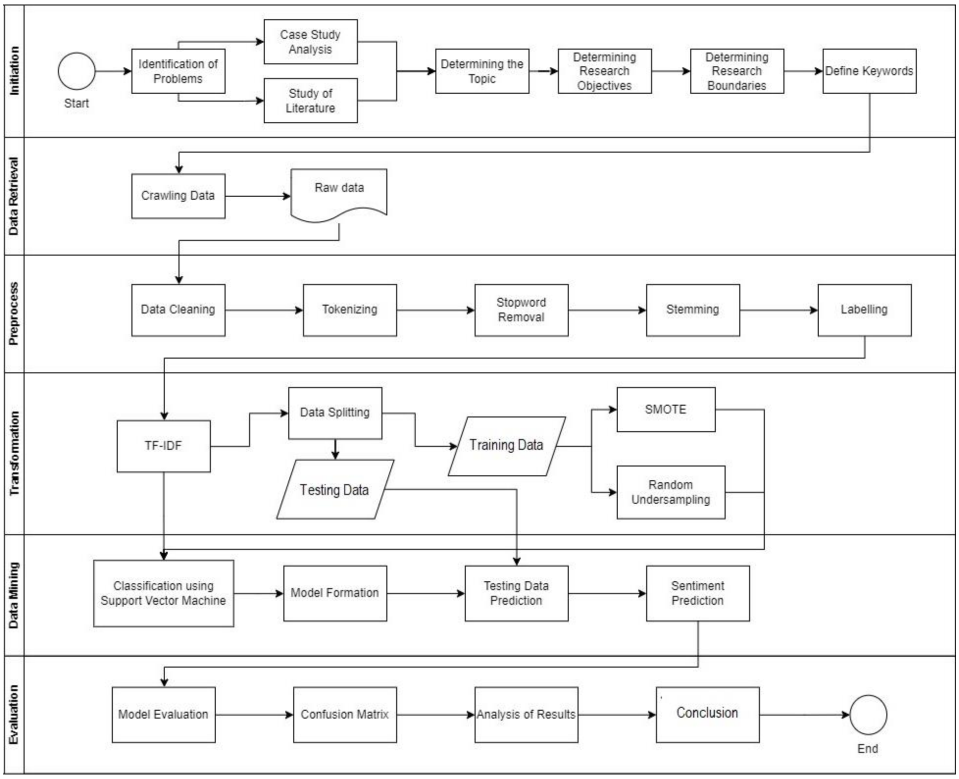


Figure 4. Systematic Problem Solving

2.1 *Scope Analysis*

At this stage, researchers will conduct sentiment analysis on case studies regarding electric vehicles, which are widely discussed on social media, especially Twitter. On this platform, there is much discussion about electric vehicles because of their characteristics of not using fossil fuels, which are thought to reduce the greenhouse effect. However, in Indonesia, there are differences of opinion among the public regarding electric vehicles, so controversy arises between supporters and opponents. Users on Twitter expressed their views using various keywords, one of which was "Electric Vehicles." Thus, researchers are interested in analyzing public responses on Twitter to look for positive sentiments for those who support the distribution of electric vehicles, as well as negative sentiments for those who disagree.

2.2 *Data Preprocessing Stage*

In this research, comments from Twitter social media users are the data source. Collecting comment data from Twitter social media through web scraping techniques using the snscrape library. The initial step in web scraping is to enter the desired keywords, "electric vehicles," in this study and determine the relevant time range and language to the comments. After that, comment data that meets these criteria is collected into a data frame using the Python programming language. This data frame contains attributes such as tweet date, username, and the content of the comments that have been collected. The dataset used for this research consists of 19530 comment data.

Next is the data preprocessing stage, which aims to prepare the unstructured text data structure to become more structured and to prepare the data before weighting and classification using the Support Vector Machine classification method. The stages that will be carried out consist of the first subprocess case folding. In the second subprocess, data cleansing is carried out; then, in the third subprocess, stopwords are removed; in the fourth subprocess, tokenizing is carried out; and in the last subprocess, the stemming process is carried out. After completing all data preprocessing steps, the next step is to conduct a manual labeling process, where each comment data will be labeled whether it is in the positive or negative class. The following in Table 1 are presented some examples of the results of data labeling that has been done.

Table 1. The Results of The Labeling Process

Tweets	Label
['transportasi', pake', listrik', bensin', tetep', aja', macet', nama', aja', kendara', pribadi',bener', bikin', transportasi', umum', nyaman']	Negatif
[pakai’, kendara’, listrik’, jamin’, keren’, banget’, hemat, no’, antri’, bensin’]	Positif
['indonesia',keren',ganti',kendara',listrik']	Positif

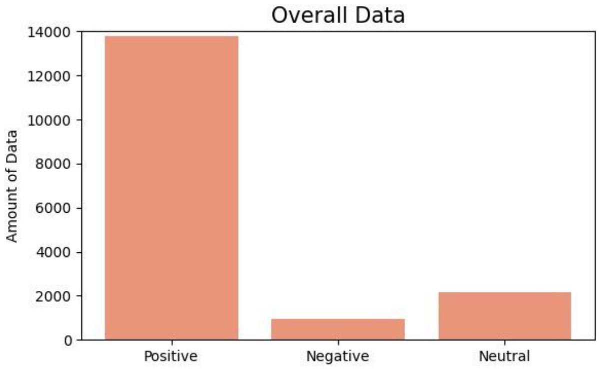


Figure 5. Amount of Data Based on Label

Based on Figure 5, the entire data has 16886 data comments. The distribution of positive and negative labels comprised 13780 positive labeled data, 931 negative labeled data, and 2175 neutral labeled data. However, the commentary data used in this study were only those labeled 'positive' and 'negative,' so data labeled 'neutral' was immediately discarded. Then, the data vectorization process is carried out to change the commentary data in the text into a numerical vector representation using the TF-IDF method. The equations used to carry out the TF-IDF process are stated in Equations (2), (3), and (4).

$$TF_{t,d} = n_{t,d} \quad (2)$$

$$IDF_t = \ln \frac{1+D}{1+n_{t,d}} + 1 \quad (3)$$

$$w_{t,d} = TF_{t,d} \times IDF_t \quad (4)$$

This research implements the three scenarios using the baseline SVM model (without imbalance handling techniques), the SMOTE method, and the RUS method. The following results in dividing the data with ratios of 70:30, 80:20, and 90:10 are shown in Table 2.

Table 2. Amount of Training Data and Testing Data

Ratio	Training Data	Testing Data	Amount
90:10	13239	1472	14711
80:20	11768	2943	14711
70:30	10297	4414	14711

After that, Figure 6 presents the results of implementing SMOTE, and RUS for balancing the training data using a data sharing ratio of 70:30.

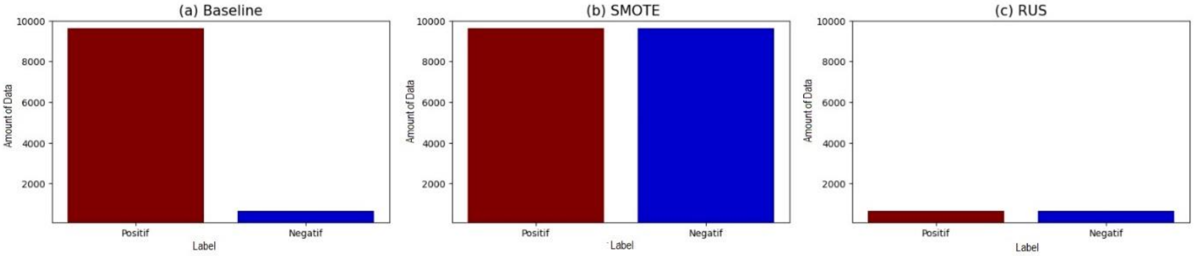


Figure 6. Comparison of Positive and Negative Charts Before and After Imbalance Handling

After the data is balanced, a classification prediction model is built using the SVM algorithm with a linear kernel. The linear kernel is a kernel whose job is to create a hyperplane with a maximum margin that separates the two classes and limits margins to have the same distance.

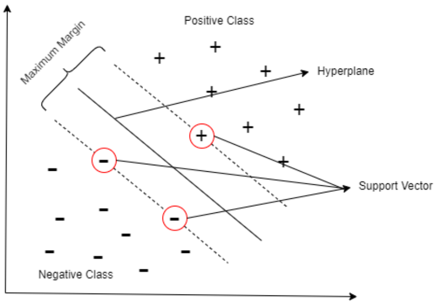


Figure 7. Simple Concept of SVM

Figure 7 shows two labels or classes: the positive class marked with a (+) sign and the positive class marked with a (-) sign. A hyperplane line separates the two labels called the decision boundary. The formula used in obtaining the hyperplane is formulated in Equation 5.

$$\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b = 0 \quad (5)$$

The data x belongs to the negative class and can be formulated in Equation 6.

$$\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b \leq 1 \tag{6}$$

The data x belongs to the positive class and can be formulated in Equation 7.

$$\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x} + b \geq 1 \tag{7}$$

In addition, there is code to carry out the classification process using data from the xtest variable, which produces new labels that can contain "Positive" or "Negative" values due to classification by the SVM model. After the classification stage, the next step is to evaluate the model's performance against the classification results. The following is a pseudocode for making equations to get the hyperplane SVM algorithm with a linear kernel.

1. Determine the data points: $\mathbf{X_i} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$, the attribute used, namely TF-IDF.
2. Determine the data class: $y_i \in \{-1, +1\}$, where y = -1 is the negative category and y = +1 is the positive category.
3. Pairing data and class $\{(x_i, y_i)\}_{i=1}^{N}$, where N is the number of data.
4. Calculate the Lagrange coefficient to maximize the margin between the hyperplane and the closest data using Equation (8).

$$Ld = \sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \sum_{i=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \tag{8}$$

Condition 0≤a_i≤C and $\sum_i^N a_i y_i = 0$. The a_i value is the value of the Lagrange Multiplier, which has a zero or positive value. The a_j value is the value of the Lagrange Multiplier, which has a value of one or negative.

5. Calculate the w value using Equation (9) and the b value using Equation (10).

$$\mathbf{w} = \sum_{i=1}^{N} a_i y_i x_i \tag{9}$$

$$b = -\frac{1}{2}\left(\mathbf{w} \cdot \mathbf{x}^{+} + \mathbf{w} \cdot \mathbf{x}^{-}\right) \tag{10}$$

where w is the weight parameter while b is the bias value. $\mathbf{x}^{+}$ is training data located at y = +1 and $\mathbf{x}^{-}$ is training data located at y = -1.

6. Create a classification decision function sign(f(x)) using the Equation (11).

$$f(\mathbf{x}) = sign[\mathbf{w} \cdot \mathbf{x} + b] \tag{11}$$

The model creation and training process is complete based on the pseudocode above. It can be continued towards the data classification process and evaluation of the classification model using test data.

2.3 Result and Conclusion Stage

The author uses the confusion matrix to evaluate the research after building a classification prediction model using the SVM algorithm with a linear kernel and carrying out the data balancing process using the SMOTE and RUS methods. In this case, the author can calculate the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values, which will then be used to produce the accuracy, recall, precision, and F1-Score. The equations for calculating accuracy, recall, precision, and F1-Score values can be found in Equations (12), (13), (14), and (15) as follows.

1. Accuracy

$$\text{accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{12}$$

2. Precission

$$\text{precission}=\frac{\text{TP}}{\text{TP+FP}} \tag{13}$$

3. Recall

$$\text{recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{14}$$

4. F1-Score

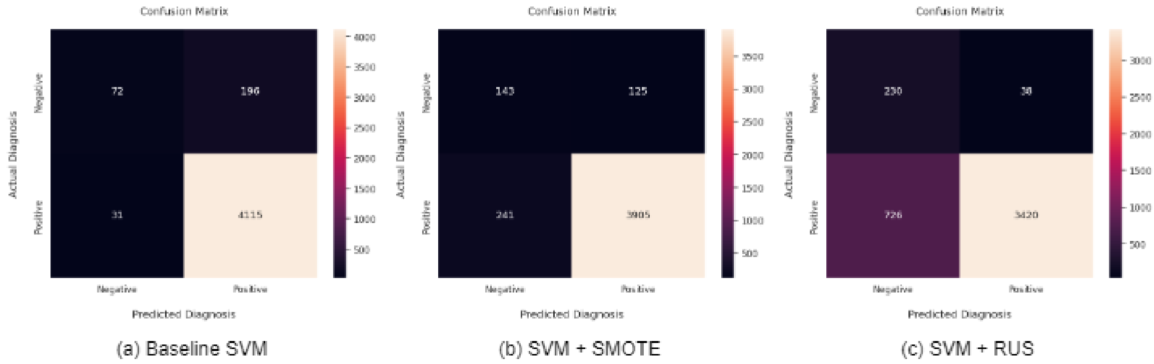
$$\text{f1-score}=\frac{2\times\text{precission}\times\text{recall}}{\text{precission+recall}} \tag{15}$$

3. Result and Discussion

Based on the training data resulting from the TF-IDF process stored in the variable x_i and the process of creating and training the SVM algorithm model using the training data that has been done before, we get a weight vector w, which has dimensions of 8,391, or it can be said $\mathbf{w} \in \mathbb{R}^{8391}$. Besides that, the value of *b* is obtained, which is the bias value. Thus, the optimal hyperplane equation is obtained, which separates the data into two classes. As an example, the hyperplane equation for the classification results with the SVM method and 70:30 data splitting is presented in Equation (16):

$$\begin{pmatrix} -0,35289 \\ 0,53072 \\ 0,14733 \\ \vdots \\ 0,13292 \end{pmatrix}^T \cdot \mathbf{x} + 0,78016 = 0 \tag{16}$$

The classification model performance evaluation technique uses the confusion matrix to identify and understand the various components of the model classification results. In the confusion matrix, there are several essential components, such as True Positive (TP), which shows the number of correct predictions from positive labels to actual values that are also positive. In addition, there is a False Positive (FP), which describes the number of incorrect predictions of positive labels to actual values that are negative. Then, there is a True Negative (TN), which reflects the number of correct predictions from a negative label to an actual value, which is also negative. Finally, the False Negative (FN) indicates the number of wrong predictions from a negative label to the true positive value. In the figure, the results of comparing the performance evaluation of the SVM algorithm classification model are presented using the confusion matrix before and after applying imbalance handling at a ratio of 70:30.



(a) Baseline SVM (b) SVM + SMOTE (c) SVM + RUS

Figure 8. Confusion Matrix Results At 70:30 Ratio

There are four components in the confusion matrix in Figure 8, namely True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP). Figure 8a presents the confusion matrix for the SVM baseline model with a ratio of 70:30. The value of TP is 4115 comments, FP is 196 comments, TN is 72 data comments, and FN is 31 data comments. Figure 8b presents the confusion matrix for the SVM baseline model that applies SMOTE as an imbalance handling technique with a ratio of 70:30. The value of TP is 3905 comments, FP is 125 data comments, TN is 143 data comments, FN is 241 data comments. Figure 8c presents the confusion matrix for the SVM baseline model that applies RUS as an imbalance handling technique with a ratio of 70:30, with a TP value of 3420 comment data, FP 38 data comments, TN 230 data comments, and FN 726 data comments.

3.1 Model Performance Evaluation Using Evaluation Metrics

Evaluation metrics are measures or numbers used to measure the performance or effectiveness of a model, algorithm, or system in completing a particular task. These metrics provide quantitative information about the extent to which the results produced correspond to predetermined goals. Evaluation metrics include accuracy, precision, recall, F1-score, and many others, which suit different problem types and analysis purposes.

Table 3 Comparison of Evaluation Metrics

Method	Ratio	Accuracy	Precision	Recall	F1-Score
SVM	70:30	94,8%	95,4%	99,2%	97,3%
	80:20	94,8%	95,5%	99,2%	97,2%
	90:10	94,9%	95,7%	99,1%	97,1%
SVM+SMOTE	70:30	91,7%	96,9%	94,1%	95,5%
	80:20	91,6%	97,2%	93,8%	95,5%
	90:10	91,1%	97%	93,4%	95,2%
SVM+RUS	70:30	82,7%	98,9%	82,5%	89,9%
	80:20	82,4%	98,9%	82,2%	89,9%
	90:10	82,1%	98,9%	81,9%	89,6%

Based on Table 3 above, the baseline SVM model obtained the best accuracy value with a ratio of 70:30 without using imbalance handling, with an accuracy of 94,8%. However, it is essential to note that the data used in this study needs to be more balanced. Previous research by Wongvorachan et al. (2023) stated that in the case of unbalanced data, the baseline SVM model tends to ignore the minority class (in this case, the "Positive" class) because it learns more from the majority class, which has a more significant amount of data. As a result, the baseline SVM model can only identify a few minority classes because their numbers are smaller.

Since the accuracy value does not fully represent the model performance in the case of imbalanced data, other evaluation metrics such as precision, recall, and F1-score were used in this study. These metrics can evaluate the performance of SVM models with various resampling techniques and data split ratios. The evaluation results show that the SVM model produces the most optimal performance if data imbalance occurs without applying a resampling approach with a ratio of 70:30.

3.2 Word Cloud Visualization

In this study, data visualization was carried out using a word cloud to display the words with the highest frequency in the comment data labeled "Positive" and "Negative." The word cloud highlights words with a larger font size that reflects a higher frequency. Figure 9a presents a word cloud in the commentary data labeled "Positive." The words "vehicle" and "electricity" have the highest frequency. The condition shows that the dominant topics and keywords in the comments are about the use of electric vehicles. In addition, the word cloud also offers many words with positive connotations, such as "environmentally friendly," which may be related to "environmentally friendly." Other terms, such as "really cool" and "healthy air," show support for using electric vehicles because of their positive impression on society and their contribution to reducing urban air pollution.

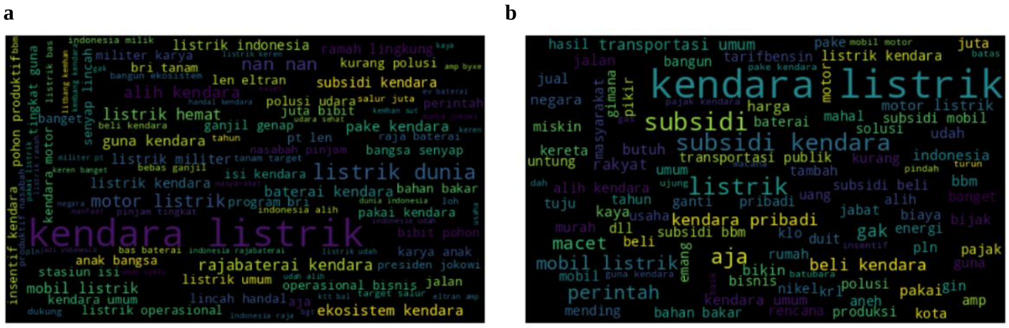


Figure 9 Word Cloud Comment Data with (a) 'Positive' Label (b) ‘Negative’ Label

Figure 9b presents a word cloud of comment data labeled "negative." The words with the most frequency are "vehicle" and "electricity." The terms are also because the entire commentary data marked "negative" has a topic, and the keywords being searched for are about the use of electric vehicles. Other words that have negative connotations include "jammed," expensive," and "strange." The number of "jammed" comments is due to government policies regarding electric vehicle subsidies that make people flock to buy electric vehicles and impact increasing congestion. Then, the "expensive" statement refers to the significant increase in household electricity costs. The "strange" word refers to people who are not used to electric vehicles and have a stigma that electric vehicles are not durable.

4. Conclusion

In this research, the Support Vector Machine algorithm was used to evaluate sentiment regarding the use of electric vehicles based on comments on the social media platform Twitter. Knowledge Discovery in Database (KDD) is an applied methodology that includes steps such as initiation, data retrieval, text pre-processing, data transformation, data mining, and evaluation. After the text pre-processing stage, 14,711 comment data were labeled. From this data, there are 13,780 data with positive labels and 931 data with negative labels. Based on the analysis of positive words in the word cloud, terms such as "environmentally friendly" and "healthy air" show that people respond positively to electric vehicles because of their environmental benefits.

On the other hand, from the analysis of negative words in the word cloud, some expressions refer to negative impacts, such as "expensive" and "strange." The terms indicate that some people may not be interested due to financial reasons and a lack of awareness about electric vehicles among the public. Evaluation of the performance of the Support Vector Machine algorithm in analyzing sentiment is carried out using a confusion matrix. The best results were found in the Support Vector Machine scenario with a 70:30 data split, resulting in an accuracy of 94.8%, precision of 95.5%, recall of 99.1%, and F-1 Score of 97.2%.

REFERENCES

[1] Potoglou, D., Song, R., & Santos, G. (2023). Public charging choices of electric vehicle users: A review and conceptual framework. Transportation Research Part D: Transport and Environment, vol. 121, no. 103824, pp. 1-22.

[2] Samuel, Y., Delima, R., & Rahmat, R. (2015). Implementasi metode k-nearest neighbor dengan decision rule untuk klasifikasi subtopik berita, *Jurnal Khatulistiwa Informatika*, vol. 10, pp. 1-14.

[3] Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings Of The 2nd International Conference on Knowledge Capture,* pp. 70-77.

[4] Pertiwi, S. R. G. (2018). Perbandingan metode k-nearest neighbor dan support vector machine dalam analisis sentimen twitter terhadap stasiun televisi berita Indonesia, [*Dissertation*], Yogyakarta: Universitas Gadjah Mada.

[5] Astuti, I. N. F., Darmawan, I., & Pramesti, D. (2020). Analisis sentimen pada data kuesioner evaluasi dosen oleh mahasiswa (edom) Prodi Sistem Informasi Telkom University menggunakan algoritma support vector machine. *eProceedings of Engineering*, vol. 7, no. 2, pp. 7018-7025.

[6] Ridwansyah, T. (2022). Implementasi text mining terhadap analisis sentimen masyarakat dunia di twitter terhadap Kota Medan menggunakan k-fold cross validation dan naïve bayes classifier. *Klik: Kajian Ilmiah Informatika dan Komputer*, vol. 2, no. 5, pp. 178-185.

[7] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and. Techniques*, Waltham: Morgan Kaufmann Publishers.

[8] Maddison, J., & Jeske, D. (2014). Fear and perceived likelihood of victimization in traditional and cyber settings. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, vol. 4, no. 4, pp. 23-40.

[9] Yulian, E. (2018). Text mining dengan k-means clustering pada tema LGBT dalam arsip tweet masyarakat Kota Bandung. *Jurnal Matematika "MANTIK*, vol. 4, no. 1, pp. 53-58.

[10] Aditya, B. R. (2015). Penggunaan web crawler untuk menghimpun tweets dengan metode pre-processing text mining. *Jurnal Infotel*, vol. 7, no. 2, pp. 93-100.

[11] Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). *Text Mining for Central Banks*. England: Center for Central Banking Studies, Bank of England.

[12] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pp. 30-38.

[13] Coletta, L. F., da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. (2014, October). Combining classification and clustering for tweet sentiment analysis. In *IEEE: 2014 Brazilian conference on intelligent systems*, pp. 210-215.

[14] Novantirani, A., Sabariah, M. K., & Effendy, V. (2015). Analisis sentimen pada twitter untuk mengenai penggunaan transportasi umum darat dalam kota dengan metode support vector machine. *eProceedings of Engineering*, vol. 2, no. 1, pp. 1177-1183.

[15] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, dan J. Caro, "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," dalam *IISA 2013*, IEEE, Jul 2013, hlm. 1–6. doi: 10.1109/IISA.2013.6623713.

[16] Chakraborty, K., Bhattacharyya, S., Bag, R., & Hassanien, A. A. (2018). Sentiment analysis on a set of movie reviews using deep learning techniques. *Social network analytics: Computational research methods and techniques*, *127*. Cambridge: Elsevier Inc.

[17] Lidya, S. K., Sitompul, O. S., & Efendi, S. (2015). Sentiment analysis pada teks Bahasa Indonesia menggunakan support vector machine (SVM) dan K-Nearest Neighbor (K-NN). *Proceeding Sentika 2015*, pp. 1-8.

[18] Goh, R. Y., & Lee, L. S. (2019). Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, no. 1974794, pp. 1-30.

[19] Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, 283-289.

[20] Mansourifar, H., & Shi, W. (2020). Deep synthetic minority over-sampling technique. *arXiv preprint arXiv:2003.09788*.

[21] Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: density-based synthetic minority over-sampling technique. *Applied Intelligence*, vol. 36, pp. 664-684.

[22] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, vol. 61, pp. 863-905.

[23] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, vo. 16, pp. 321-357.

[24] Hennenfent, G., & Herrmann, F. J. (2008). Simply denoise: Wavefield reconstruction via jittered undersampling. *Geophysics*, vol. 73, no. 3, pp. 19-28.

[25] Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015, August). Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE international conference on information reuse and integration*, pp. 197-202.

[26] Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, vol. 14, no. 54, pp. 1-15.

[27] Irawaty, I., Andreswari, R., & Pramesti, D. (2020, September). Vectorizer comparison for sentiment analysis on social media youtube: A case study. In *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, pp. 69-74.

[28] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., & Amirat, Y. (2015). Physical human activity recognition using wearable sensors. *Sensors*, vol. 15, no. 12, pp. 31314-31338.

[29] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *Maics*, vol. 710, no. 1, pp. 120-127.

[30] Nasution, M. R. A., & Hayaty, M. (2019). Perbandingan akurasi dan waktu proses algoritma K-NN dan SVM dalam analisis sentimen twitter. *J. Inform*, vol. 6, no. 2, pp. 226-235.