Scientific Journal of Informatics
Vol. 11, No. 2, May 2024

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Classification Model of Public Sentiments About Electric Cars Using Machine Learning

Nurul Romadoni[1*], Amril Mutoi Siregar[2], Dwi Sulistya Kusumaningrum[3], Tatang Rohana[4]
[1,2,3,4]Department of Informatics, Faculty of Computer Science, Universitas Buana Perjuangan Karawang, Indonesia

Abstract.
Purpose: This research compared the accuracy level of six algorithms based on the ROC method and the Confusion Matrix evaluation on data regarding public sentiments towards electric cars.
Methods: Data collection was conducted for data sourced from TikTok. Next, the data underwent text preprocessing (data cleaning and case folding) and text processing (stemming, tokenizing, stopword removal, word frequency, word relation, TF-IDF, scoring, and labeling). Modeling was then conducted using supervised (labeled) algorithms consisting of the Support Vector Machine (SVM), Decision Tree, Naive Bayes, Random Forest, K-Neighbor, and Logistic Regression. Finally, an evaluation was conducted (confusion matrix and ROC).
Result: The results revealed that the Decision Tree algorithm with the Confusion Matrix and ROC evaluation obtained the highest result of 87%. The algorithm with the lowest result is KNN, which has an accuracy of 56%. The classification result for the neutral sentiment has a percentage of 57.1%, followed by negative sentiment at 26.8% and positive sentiment at 16.1%. The KNN algorithm is suitable for large and low-dimensional data, SVM is suitable for data with many features and clear separation between classes, and Naive Bayes is efficient for large datasets with many low-quality features. Additionally, the Random Forest algorithm could overcome overfitting and unbalanced data. Logistic regression is also suitable for linear data without assuming a certain distribution. The Decision Tree algorithm is good for complex data as it provides a visual explanation of predictions. In this study, the Decision Tree algorithm obtained high results because it has the best characteristics and is a linear technique.
Novelty: This study found that based on the ROC method and the Confusion Matrix evaluation conducted, the Decision Tree algorithm is more accurate than the other algorithms studied.

Keywords: Classification, Electric car, Machine learning, Sentiment analysis
Received February 2024 / **Revised** April 2024 / **Accepted** May 2024

INTRODUCTION

Indonesia is increasingly committed to developing sustainable transportation by presenting electric vehicles (EVs) as an environmentally friendly solution [1]. Government programs have also been increasing their support for EVs in Indonesia to reduce greenhouse gas emissions and the country's dependence on fossil fuels. The development of EVs is expected to significantly increase the population of electric vehicles in Indonesia, which will have a positive impact on reducing air pollution and improving air quality in big cities like Jakarta. The government supports the electric vehicle program in Indonesia through Presidential Instruction Number 7 of 2022 concerning The Use of Battery-based Electric Motor Vehicles (battery electric vehicles) as Operational Service Vehicles or Individual Service Vehicles for Central Government Agencies and Local Governments [2]. However, electric vehicles also have disadvantages, such as the time required to charge a car battery, which takes a long time. In addition, the cost required to buy an electric vehicle is greater than that of conventional vehicles [3]. Thus, the outcomes of the reviews processed in this study can provide input for automotive companies in adjusting their vehicles to consumer needs and desires. This insight is important because the public's interest in electric vehicles is increasing with the availability of electric vehicles [4].

Sentiment analysis is a process that can be utilized to extract an opinion or review published on social media, such as TikTok, Twitter, WhatsApp, Instagram, YouTube, and others [5]. This research used the

*Corresponding author.
Email addresses: if20.nurulromadoni@mhs.ubpkarawang.ac.id (Romadoni), amrilmutoi@ubpkarawang.ac.id (Siregar)
DOI: 10.15294/sji.v11i2.1309

TikTok platform to conduct a sentiment analysis on digital texts about electric vehicles. TikTok is one of the fastest-growing social media platforms in the world [6]. The platform was first launched in 2017, allowing users to create and share short videos and becoming the most downloaded non-gaming application platform worldwide [7]. The TikTok platform has significantly influenced the introduction and popularization of electric cars (EVs). Various electric car brands use TikTok as a marketing platform to reach out to the younger generation active on the platform. In this way, TikTok helps raise consumer awareness and interest in EVs.

Sentiment analysis cases can be solved using the following algorithms: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Naïve Bayes (NB). Research conducted by [8] used data containing reviews of the TikTok application on Google Playstore. The methods used are VADER sentiment to label reviews, Random Under-sampling (RUS), Random Over-sampling (ROS) methods to balance data, and the Support Vector Machine algorithm. Good results are obtained using the Support Vector Machine with an F1-Score of 0.80. Furthermore, research comparing the Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor algorithms in conducting sentiment analysis of the 2019 presidential candidates of the Republic of Indonesia collected from the Twitter application was conducted by [9]. The highest accuracy results were obtained from the Naïve Bayes method at 80.90%. Meanwhile, the K-Nearest Neighbor produced an accuracy of 75.58% and the SVM method obtained the lowest accuracy at 63.99%. Moreover, research conducted by [10] predicted tourist reviews using the Naïve Bayes Classifier. They used review data for Jimbaran Beach and Kuta Beach. The accuracy results obtained are 80% for the Jimbaran Beach review data and 64% for the Kuta Beach review data. Sentiment analysis on Samsung phone review data from YouTube has also been conducted by comparing Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Naïve Bayes, and Random Forest algorithms. Logistic Regression obtained the greatest accuracy results at 90% [11]. In addition, [12] conducted a sentiment analysis on the topic of electric cars using Support Vector Machine (SVM) and Naive Bayes. SVM obtained the highest accuracy results at 90%.

Based on the problems described, this study aims to compare the accuracy level of six algorithms for review data about electric vehicles on the TikTok platform. The algorithms used are Support Vector Machine (SVM), Decision Tree, Naive Bayes, Random Forest,
K-Nearest Neighbor, and Logistic Regression.

METHODS

Several The research stages conducted are depicted in Figure 1.

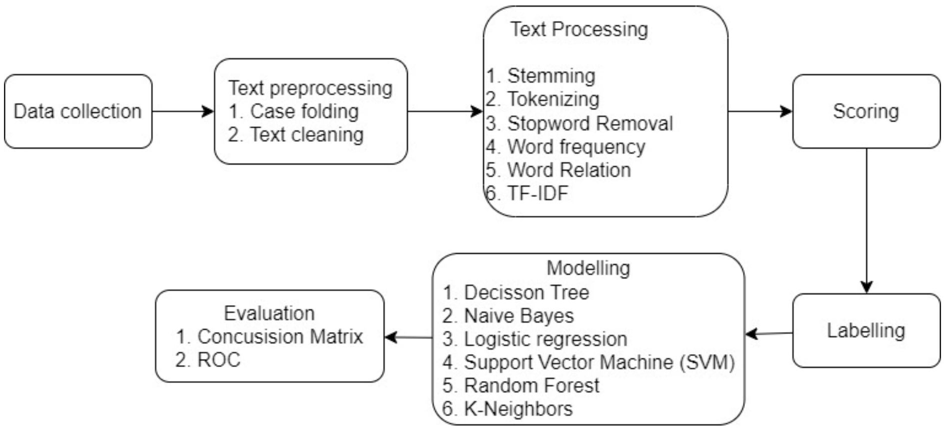

Figure 1. Research method

Data collection

The first stage of this research is data collection. The dataset is primary data from TikTok with the keyword "electric car." Data retrieval was conducted using crawling techniques such as Instant Data Scrapper. The amount of data collected is 4759 lines with two attributes: "no" and "Text." Next, the collected data underwent Text Preprocessing through Case Folding and Text Cleaning. Followed by text processing using stemming, tokenizing, stopword removal, word frequency, word relation, and TF-IDF. Then, the data underwent the scoring and labeling process, followed by the Machine Learning modeling process consisting of 6 algorithms: Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, and K-Neighbor. The implemented algorithms are then evaluated using the Confusion Matrix and Receiver Operating Curve (ROC) methods.

Text preprocessing

Text Preprocessing is a process that transforms unstructured data into data ready for processing [13]. The first stage of the Text Preprocessing conducted in this study is Case Folding and Text Cleaning. Case Folding changes all letters originally capitalized into lowercase letters, whereas Text Cleaning cleans the text from numbers, URLs, symbols, punctuation marks, corrupted data, empty rows, and columns [14]. Generally, text preprocessing stages include case folding, tokenization, text cleaning, stopword removal, stemming or lemmatization, and normalization.

Text processing

The Text Preprocessing stage is followed by the Text Processing stage. The following are the Text Processing stages used in this research:

1) Stemming is the process of finding the basis of each word so that it is formed into its basic word (stem). The stemming process is useful for improving model performance by reducing time and space complexity so that recall will increase and decrease precision [15]. In sentiment analysis, words that share the same root but have different affixes can be classified uniformly. For example, the words "like" and "really like" can be simplified to the root word "like." By applying stemming, sentiment analysis can be more accurate as it considers the similarity of word meanings of the same root word [15].
2) The tokenization process breaks down sentences into units, such as words or tokens [16]. In sentiment analysis, tokenization allows researchers to analyze each word separately and consider its meaning and sentiment. By using tokenization, sentiment analysis can be more efficient and accurate by focusing on each word individually.
3) Stopword Removal consists of the deletion of words that are not important or meaningless [17]. Words that are removed are usually of the affix conjunction type. Stopwords are useful for reducing the amount of text by 35-45%, by reducing the amount of text, the size of the dataset decreases, and the time to train the model decreases [17]. Stopword removal can improve sentiment analysis by reducing noise, speeding up analysis, and increasing accuracy. Removing stopwords reduces common words with no special meaning, reduces computational complexity, and makes important words more prominent [15].
4) TF-IDF, a statistical method algorithm to assess how important a word is [18]. TF-IDF (Term Frequency-Inverse Document Frequency) is very effective in sentiment analysis as it assigns weights to words based on how often they appear in a particular document (TF) and how common they are across documents (IDF). TF-IDF helps identify keywords that might determine positive or negative sentiments, thus improving the accuracy of sentiment analysis [19]. The following is the TF-IDF weighting scheme:

$$TFIDF_J = TF \times IDF \quad (1)$$

Description

$TF\ (t, d)$: The frequency of words appearing in the document.

$$TF = \frac{\text{the number of occurrences of a word in a document d}}{\text{number of words in document d}} \quad (2)$$

$IDF\ (t, D)$: calculates how unique a word is in the entire document.

$$IDF = \frac{\text{number of documents in collection D}}{\text{the number of documents in collection D that contain the word } t+1} \quad (3)$$

5) Scoring in sentiment analysis involves assigning values or scores to data that describe positive, negative, or neutral sentiments. This step is important for understanding user feedback, decision- making, competitive analysis, and improving customer satisfaction. Scoring helps analyze data quickly and efficiently [20].
6) Labeling in sentiment analysis is the process of assigning a particular label or category to a text or document based on its sentiment, such as positive, negative, or neutral. This process can be done manually by humans or by using algorithms to identify sentiment from text. Labeling is important because it makes it possible to analyze sentiment systematically and can help in decision-making [21].

Logistic Regression (LR)

Logistic Regression has similarities with Linear Regression when using binominal response variables [20]. This algorithm models the relationship between a binary dependent variable (two categories) and one or more independent variables. Logistic regression is suitable for sentiment analysis due to its easy interpretation, computational efficiency, and ability to handle linear relationships. Regularization prevents overfitting [23]. The logistic function (sigmoid function) transforms continuous values into a range between 0 and 1. The logistic curve or logistic function allows non-linearity to be applied [21].

$$1 = log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_0 + \beta_2 x_2 \ ... \ \beta_0 + \beta_n x_n \quad (4)$$

Description:

p : probability estimation

β_0 : intercept y

β_1 : is the variable coefficient x_1

β_2 : is the variable coefficient x_2

β_n : is the variable coefficient x_n

Support Vector Machine (SVM)

SVM is a multifunctional algorithm that can be used for regression and classification problems with minor modifications [22]. SVMs have been widely used and proven successful in sentiment analysis and text-mining tasks. One of the features of supervised learning methods is the ability of SVMs to solve classification problems using kernels, where the data is transformed to a larger dimensional space to be classified as non-linearly separable data. Previous studies mentioned that the SVM algorithm proved to be a good classification algorithm by producing 98% accuracy [20]. The following is the Support Vector Machine (SVM) algorithm equation.

$$\{(x_{1,}y_1), (x_{2,}y_2), (x_{3,}y_3), (x_{4,}y_4), \dots , \quad (x_n \, y_n)\} \quad (5)$$

In this equation, the constant yynn = -1 / -1 indicates the position of point XXnn in nn with value n, which is the number of samples, and each XXnn is a vector of real values with dimension p. To maintain a feature or variable with a larger variance value, the determination of the scale becomes very important to find the required hyperplane.

$$w.x + b = 0 \quad (6)$$

The hyperplane will go out of bounds if there is no parameter b, resulting in a finite solution.

K-Nearest Neighbors (KNN)

K-Nearest Neighbor, also known as KNN, is a machine learning algorithm often used in solving classification or regression problems. The algorithm will predict data points by utilizing the Kth data point closest to the average or selecting the majority class. Each data point in the training data will be compared, and the closest data point will be used. However, KNN is very sensitive to distance measure and parameter selection. The most common distance measure is Euclidean [23]. A previous study also mentioned that the KNN algorithm is the best machine learning algorithm as it obtained an accuracy result of 84.93% [30].

Random Forest (RF)

Random Forest (RF) is an ensemble learning algorithm useful for handling classification and regression cases. The RF algorithm can predict a set of random regression trees [24]. The performance of RF is determined by the number of trees that contain the predictor. The method to determine how many trees are needed is to compare the predictions made by the forest and those made by a subset of the forest [25].

Decision Tree (DT)

Decision Tree techniques are very effective in creating classification models with available data [26]. Previous research mentioned that the DT algorithm obtained the highest accuracy results with 85% accuracy results [30]. By using a predictive model that resembles a tree or hierarchical structure, the decision tree algorithm extracts knowledge from data, and then the decision tree idea transforms the data into decision rules and hierarchies. A decision tree can be interpreted as a collection of hyperplanes, where each branch points to one of the axes. Generally, the complexity of a tree can be measured by the total number of nodes, the number of leaves, the depth of the tree, and the number of attributes used. The following is the decision tree entropy equation:

$$E(D) = \sum_{i=1}^{n} -p_{C(i)}(p_{C(i)}) \quad (7)$$

Description:

$p_{C(i)}$: probability of class c (i) in a node.

Entropy D : a measure of the irregularity of the considered sample.

Naïve Bayes (NB)

The Naïve Bayes algorithm is a collection of classification algorithms based on the Bayes Theorem. This algorithm works by calculating the frequency and combination of values in a set group [27]. Naïve Bayes can be used for classification cases in sentiment analysis and can provide a strong hypothesis of any conditions and events. The probability calculation for Naïve Bayes is shown below.

$$P\ (m|n) = \frac{\mathrm{P(n|m)\ P(m)}}{\mathrm{P(n)}} \quad (8)$$

Description:

P (m|n) : probability of class x. x is the target, and the prediction is the attribute,

P (m) : probability of the previous class,

P (n|m) : probability of the predictor that has been assigned a class,

P (n) : previous predictor probability.

Confusion matrix

The Confusion Matrix is used to evaluate an algorithmic model used to solve classification cases. It has columns that represent the predicted result class and the actual result class. The attributes in the Confusion Matrix consist of horizontal attributes (classified objects) and vertical attributes (actual objects) [28]. The Confusion Matrix has four terms: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). This evaluation method is used to determine the accuracy, precision, recall, and F1-Score [29]. The Confusion Matrix equation is shown below [30].

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (9)$$

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (10)$$

$$\text{Acuraccy} = \frac{\mathrm{(TP+TN)}}{\mathrm{(TP+TN+FP+FN)}} \quad (11)$$

Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic is an evaluation method to measure the performance of the algorithm used. ROC is commonly used in classification problems and calculates a probability curve. On the ROC curve, the True Positive Rate (TPR) is on the y-axis, and the False Positive Rate (FPR) is on the x-axis [31]. ROC is based on specification (True Negative Rate) and recall (True Positive Rate). Specification measures the performance of the negative section, while recall measures the performance of the positive section [32].

RESULTS AND DISCUSSIONS

Data collection

This research’s dataset was obtained through the TikTok platform from November 15 to November 18, 2023. A total of 13,326 comments with two attributes was collected (Table 1).

Table 1. Data collection results

No.	Text
	Klo tiba" habis di jln emang harus di dorong ke rumah baru di cas ??
	berarti dengan kata lain mobil listriknya kurang mahal bos 😂
	mending beli yang pasti aja mobil nya bensin aja 10k per liter sama kopi 5k beli di pinghir pombensin
	-
	-
	-
13.326	charging station antar kota masih sangat jarang, berbeda dengan negara CHINA sudah banjir charging station nya!

Text preprocessing

The next stage is the text cleaning process consisting of removing symbols, numbers, url addresses, empty lines, and duplicate data. This stage also includes case folding, which converts uppercase letters into lowercase letters.

Table 2. Text cleaning results

No.	Before	After
1.	Klo tiba" habis di jln emang harus di dorong ke rumah baru di cas ??	klo tiba habis di jln emang harus di dorong ke rumah baru di cas
2.	berarti dengan kata lain mobil listriknya kurang mahal bos 😂	berarti dengan kata lain mobil listriknya kurang mahal bos
3.	mending beli yang pasti aja mobil nya bensin aja 10k per liter sama kopi 5k beli di pinghir pombensin	mending beli yang pasti aja mobil nya bensin aja 10k per liter sama kopi 5k beli di pinghir pombensin
-	-	-
-	-	-
-	-	-
13.326	charging station antar kota masih sangat jarang, berbeda dengan negara CHINA sudah banjir charging station nya!	charging station antar kota masih sangat jarang berbeda dengan negara china sudah banjir charging station nya

Text processing

The initial stage of text processing is stemming, which replaces words with affixes into basic words.

Table 3. Text stemming

No.	Before	After
1.	klo tiba habis di jln emang harus di dorong ke rumah baru di cas	klo tiba habis di jln emang harus di dorong ke rumah baru di cas
2.	berarti dengan kata lain mobil listriknya kurang mahal bos	arti dengan kata lain mobil listrik kurang mahal bos
3.	mending beli yang pasti aja mobil nya bensin aja 10k per liter sama kopi 5k beli di pinghir pombensin	mending beli yang pasti aja mobil nya bensin aja k per liter sama kopi k beli di pinghir pombensin
-	-	-
-	-	-
-	-	-
13.326	charging station antar kota masih sangat jarang berbeda dengan negara china sudah banjir charging station nya	charging station antar kota masih sangat jarang beda dengan negara china sudah banjir charging station nya

The second stage, tokenizing, aims to break the stemming results into words. The tokenization results are shown in Table 4.

Table 4. Text tokenization

No.	Before	After
1.	klo tiba habis di jln emang harus di dorong ke rumah baru di cas	klo tiba habis jln emang harus di dorong ke rumah baru di cas

The third stage is stopword removal to eliminate words that are not meaningful. The stopword removal results are shown in Table 5.

Table 5. Stopword removal

No.	Before	After
1.	klo tiba habis jln emang harus di dorong ke rumah baru di cas	 habis dorong cas

The fourth stage is the calculation of words using Document Frequency (DF), Term Frequency (TF), Inverse Document Frequency (IDF) and Term Frequency/Inverse Document Frequency (TF/IDF). The calculation was conducted to determine the frequency of words. Table 6 shows word frequency process results.

Table 6. Word frequency

No.	Word	n
1.	mobil	1643
2.	Listrik	1401
3.	Harga	392
4.	Bensin	258
5.	Mahal	250
6.	Batre	207
7.	Cas	197
8.	ngecas	189
9.	Baterai	187
10.	Murah	177

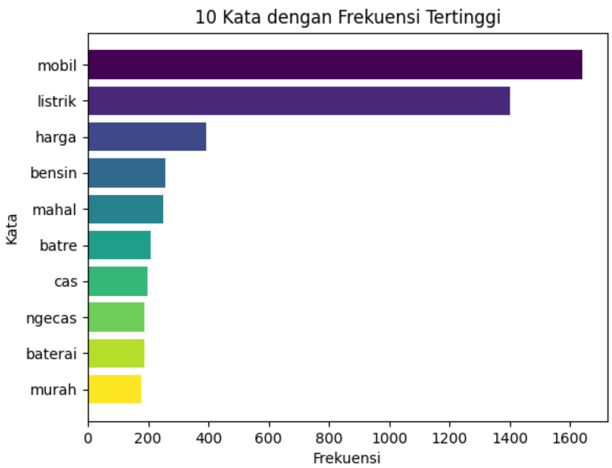

Figure 2. Word Frequency

The fifth stage is the TF-IDF method for weighting words. The TF-IDF results are shown in Table 7. The next process is to determine word-linkages and the results are shown in Table 8. Figure 2 shows the Wordcloud results.

Table 7. TF-IDF

No.	Word	n	tf	idf	Tf_idf
1.	Mobil	1643	0.000727	7.74464	0.00563
2.	Listrik	1401	0.000613	7.74464	0.00475
3.	Harga	392	0.000422	7.74464	0.00327
4.	Bensin	258	0.000144	7.74464	0.00111
5.	Mahal	250	0.000659	7.74464	0.00510
6.	Batre	207	0.000113	7.744647	0.00087
7.	Cas	197	0.000210	7.744647	0.00163
8.	ngecas	189	0.000778	7.744647	0.00602
9.	Baterai	187	0.0001087	7.744647	0.00084
10.	Murah	177	0.0007485	7.744647	0.005797

Table 8. Word linkage

No.	Word	n
1.	mobil-listrik	756
2.	Mobil-mobil	203
3.	Listrik-listrik	174
4.	Listrik-mobil	171
5.	Harga-mobil	147
6.	Mobil-bensin	122
7.	Mobil-mahal	81
8.	Listrik-harga	66
9.	Listrik-bensin	65
10.	Listrik-mahal	59

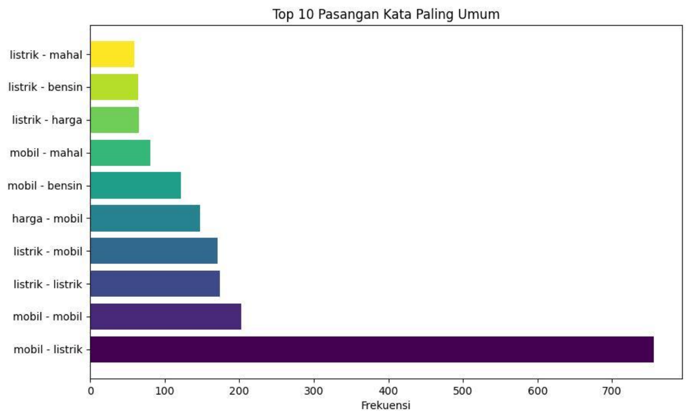

Figure 3. Word linkage

The next text processing step is to determine the sentiment score; a score of -4 to -1 indicates a negative sentiment, a score of 0 indicates a neutral sentiment and a score of 1 to 4 indicates a positive sentiment. Scoring is conducted based on positive and negative data sets on the github platform with a library of positive and negative data. The results of the sentiment score calculation on the set of sentences are shown in Table 9.

Table 9. Text scoring

No.	Score	After
1.	1	klo tiba habis di jln emang harus di dorong ke rumah baru di cas
2.	-1	arti dengan kata lain mobil listrik kurang mahal bos
3.	0	mending beli yang pasti aja mobil nya bensin aja k per liter sama kopi k beli di pinghir pombensin
-	-	-
-	-	-
-	-	-
13.326	-2	charging station antar kota masih sangat jarang beda dengan negara china sudah banjir charging station nya

Next, labels were made based on the scores obtained. The results are shown in Table 10 and Figure 4.

Table 10. Labelling

No.	Classification	After
1.	Neutral	klo tiba habis di jln emang harus di dorong ke rumah baru di cas
2.	Negative	arti dengan kata lain mobil listrik kurang mahal bos
3.	Neutral	mending beli yang pasti aja mobil nya bensin aja k per liter sama kopi k beli di pinghir pombensin
-	-	-
-	-	-
-	-	-
13.326	Negative	charging station antar kota masih sangat jarang beda dengan negara china sudah banjir charging station nya

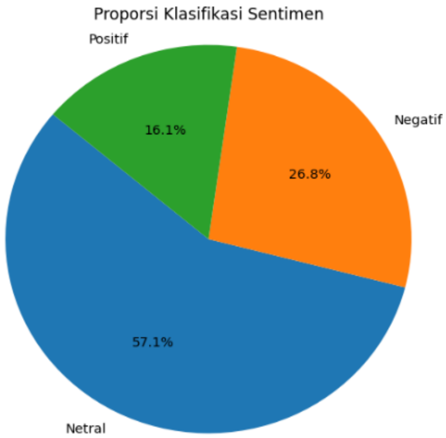

Figure 4. Proportion of classified sentiments

Figure 4 shows that there are more neutral sentiments than positive and negative, with 57.1% for neutral sentiments, 26.8% for negative sentiments, and 16.1% for positive sentiments.

Table 11. Percentage of classified sentiments

Classification of sentiments	Percentage
Negative	26.8%
Positive	16.1%
Neutral	57.1%

As shown in Table 11, 57.1% of the respondents' responses were neutral. Meanwhile, 26.8% of the responses fell into the negative category, and 16.1% were positive. These results indicate that most Indonesians tend not to have strong views on electric vehicles. Although some respondents gave negative and positive responses, the percentage was relatively lower than the neutral responses.

Comparison of algorithms

This research compared six algorithms and evaluated them using a confusion matrix and ROC. The following are the Confusion Matrix results of the six algorithms.

Table 12. Confusion Matrix Results

Algorithm	Accuracy	Precision	F1-Score	Recall
Naive Bayes	69.26%	0.68%	67%	57%
Logistic Regression	78.02	78%	77%	78%
Support Vector Machine	71.92%	73%	69%	72%
Random Forest	82.02%	82%	82%	82
Decision Tree	84.79%	85%	85%	85%
K-Neighbors	60.27%	63%	49%	60%

Figure 5 shows that the highest accuracy results from the Confusion Matrix and ROC are obtained by the Decision Tree algorithm at 0.84 and 0.87, respectively. Followed by the Random Forest algorithm with an accuracy of 0.83 and ROC of 0.81. Next, the Logistic Regression has an accuracy of 0.78 and ROC of 0.77. Meanwhile, the Support Vector Machine algorithm obtained an accuracy of 0.72 and an ROC of 0.62, and the Naïve Bayes algorithm has an accuracy and ROC of 0.69 and 0.67, respectively. The smallest accuracy and ROC results were obtained by the K-Nearest Neighbor algorithm at 0.60 and 0.55, respectively.

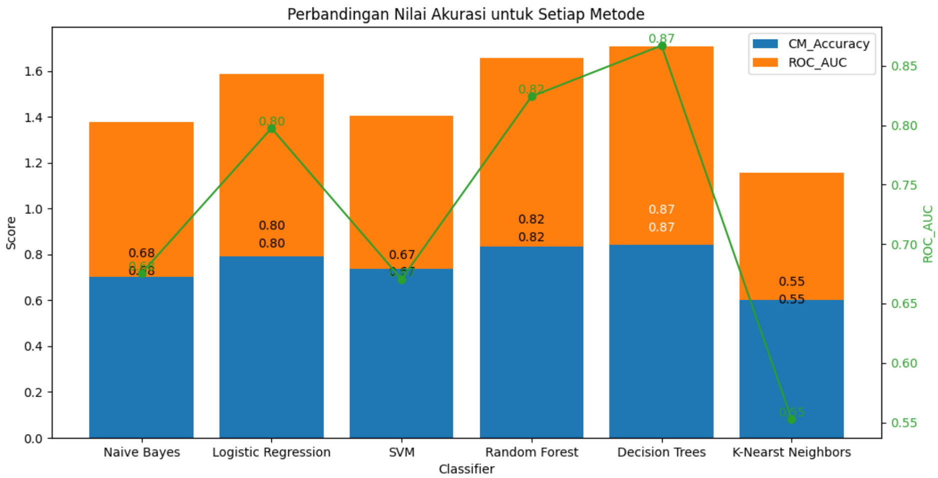

Figure 5. Model comparison results from using the Confusion Matrix and ROC.

Figure 5 shows the comparison results from using the Confusion Matrix and ROC. The Naïve Bayes algorithm obtained an accuracy result of 68%. Meanwhile, the Logistic Regression algorithm's accuracy results were 80%, 67% for the Support Vector Machine (SVM), 82% for the Random Forest, 87% for the Decision Tree, and 55% for the KNN algorithm. Thus, the Decision Tree algorithm obtained the highest confusion matrix and ROC results, while KNN obtained the lowest results.

The results show that the Decision Tree algorithm outperformed other algorithms because they tend to be easier to interpret as they produce feature-based rules that can be explained intuitively. Decision Trees do not rely on distributional assumptions that must be met by the data. This feature allows the algorithm to work well on a wide range of data types without considering whether the data distribution conforms to certain assumptions. Decision Trees can recognize relationships and interactions between various features in the dataset, allowing the algorithm to find more complex sentiment patterns on complex data.

CONCLUSION

This research used text mining to conduct sentiment classification of electric vehicle reviews by comparing the accuracy of six algorithms: Support Vector Machine (SVM), Decision Tree, Naive Bayes, Random Forest, K-Neighbor, and Logistic Regression. The classification analysis showed that neutral sentiment has the largest percentage at 57%, followed by negative sentiment at 26.8% and positive sentiment at 16.1%. Next, the accuracy rate obtained from the confusion matrix and ROC evaluation indicates that the Decision Tree algorithm had the highest result at 87%. In contrast, the KNN algorithm achieved the lowest accuracy at 56%. The difference in accuracy results is due to the ability of the Decision Tree algorithm to produce rules that are easy to interpret. Meanwhile, the KNN algorithm tends to have lower interpretation, which results in lower accuracy. Thus, the selection of algorithms in sentiment analysis greatly affects the final results and interpretation of the analysis results.

REFERENCES

[1] Direktorat Utama Pembinaan dan Pengembangan Hukum Pemeriksaan Keuangan Negara Badan Pemeriksa Keuangan, "7 Tahun 2022 Penggunaan Kendaraan Bermotor Listrik Berbasis Baterai (Battery Electric Vehicle) Sebagai Kendaraan Dinas Operasional dan/atau Kendaraan Perorangan Dinas Instansi Pemerintah Pusat dan Pemerintahan Daerah," peraturan.bpk.go.id.

[2] I. Kosmidis, D. Müller-Eie, and A. Delbosc, "Electric cars as a path to sustainable travel behaviour: Insights from Nord-Jæren," *Transp Res D Transp Environ*, vol. 125, p. 103982, Dec. 2023, doi: 10.1016/j.trd.2023.103982.

[3] V. Tulus and P. Sidabutar, “Kajian pengembangan kendaraan listrik di Indonesia: prospek dan hambatannya,” 2020.
[4] R. Jena, “An empirical case study on Indian consumers’ sentiment towards electric vehicles: A big data analytics approach,” *Industrial Marketing Management*, 2020, doi: 10.1016/j.indmarman.2019.12.012.
[5] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, “Transforming sentiment analysis in the financial domain with ChatGPT,” *Machine Learning with Applications*, vol. 14, p. 100508, Dec. 2023, doi: 10.1016/j.mlwa.2023.100508.
[6] S. Easwar, S. Alonzi, J. Hirsch, B. Mossman, and M. Hoerger, “Palliative Care TikTok: Describing the Landscape and Explaining Social Media Engagement,” *J Palliat Med*, vol. 26, no. 3, 2023, doi: 10.1089/jpm.2022.0250.
[7] C. Zhu, X. Xu, W. Zhang, J. Chen, and R. Evans, “How health communication via tik tok makes a difference: A content analysis of tik tok accounts run by Chinese provincial health committees,” *Int J Environ Res Public Health*, vol. 17, no. 1, Jan. 2020, doi: 10.3390/ijerph17010192.
[8] M. Isnan, G. N. Elwirehardja, and B. Pardamean, “Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model,” *Procedia Comput Sci*, vol. 227, pp. 168–175, 2023, doi: 10.1016/j.procs.2023.10.514.
[9] M. Wongkar and A. Angdresey, “Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter,” in *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, 2019. doi: 10.1109/ICIC47613.2019.8985884.
[10] A. V. D. Sano, A. A. Stefanus, E. D. Madyatmadja, H. Nindito, A. Purnomo, and C. P. M. Sianipar, “Proposing a visualized comparative review analysis model on tourism domain using Naïve Bayes classifier,” *Procedia Comput Sci*, vol. 227, pp. 482–489, 2023, doi: 10.1016/j.procs.2023.10.549.
[11] M. Syamala and N. J. Nalini, “A speech-based sentiment analysis using combined deep learning and language model on real-time product review,” *International Journal of Engineering Trends and Technology*, vol. 69, no. 1. 2021. doi: 10.14445/22315381/IJETT-V69I1P226.
[12] N. Wayan Ernawati, I. Nyoman Satya Kumara, W. Setiawan, J. Raya Kampus Unud Jimbaran, K. Kuta Sel, and K. Badung, “Perbandingan Metode Klasifikasi Support Vector Machine Dan Naïve Bayes Pada Analisis Sentimen Kendaraan Listrik,” 2023.
[13] A. A. Firdaus, A. Yudhana, and I. Riadi, “Public Opinion Analysis of Presidential Candidate Using Naïve Bayes Method,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 2023, doi: 10.22219/kinetik.v8i2.1686.
[14] A. W. Pradana and M. Hayaty, “The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 375–380, Oct. 2019, doi: 10.22219/kinetik.v4i4.912.
[15] A. Jabbar, S. Iqbal, M. I. Tamimy, S. Hussain, and A. Akhunzada, “Empirical evaluation and study of text stemming algorithms,” *Artif Intell Rev*, vol. 53, no. 8, pp. 5559–5588, Dec. 2020, doi: 10.1007/s10462-020-09828-3.
[16] Gunawansyah, R. Rahayu, Nurwathi, B. Sugiarto, and Gunawan, “Automated essay scoring using natural language processing and text mining method,” in *Proceeding of 14th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/TSSA51342.2020.9310845.
[17] D. J. Ladani and N. P. Desai, “International Conference on Advanced Computing and Communication Systems (ICACCS),” *h International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2020.
[18] C.-Z. Liu, Y.-X. Sheng, Z.-Q. Wei, and Y.-Q. Yang, “Research of Text Classification Based on Improved TF-IDF Algorithm,” 2018.
[19] S. Fransiska and A. Irham Gufroni, “Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method,” *Scientific Journal of*

Informatics, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji

[20] D. Muhajir, M. Akbar, A. Bagaskara, and R. Vinarti, "Improving classification algorithm on education dataset using hyperparameter tuning," in *Procedia Computer Science*, 2021. doi: 10.1016/j.procs.2021.12.171.

[21] R. Khan, F. Rustam, K. Kanwal, A. Mehmood, and G. S. Choi, "US Based COVID-19 Tweets Sentiment Analysis Using TextBlob and Supervised Machine Learning Algorithms," in *2021 International Conference on Artificial Intelligence, ICAI 2021*, 2021. doi: 10.1109/ICAI52203.2021.9445207.

[22] R. Ameri, C. C. Hsu, S. S. Band, M. Zamani, C. M. Shu, and S. Khorsandroo, "Forecasting PM 2.5 concentration based on integrating of CEEMDAN decomposition method with SVM and LSTM," *Ecotoxicol Environ Saf*, vol. 266, Nov. 2023, doi: 10.1016/j.ecoenv.2023.115572.

[23] Y. M. Wazery, E. Saber, E. H. Houssein, A. A. Ali, and E. Amer, "An Efficient Slime Mould Algorithm Combined with K-Nearest Neighbor for Medical Classification Tasks," *IEEE Access*, vol. 9, pp. 113666–113682, 2021, doi: 10.1109/ACCESS.2021.3105485.

[24] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13. Institute of Electrical and Electronics Engineers Inc., pp. 6308–6325, 2020. doi: 10.1109/JSTARS.2020.3026724.

[25] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[26] P. Sudhir and V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis," *Global Transitions Proceedings*, vol. 2, no. 2, 2021, doi: 10.1016/j.gltp.2021.08.004.

[27] M. M. Saritas and A. Yasar, "International Journal of Intelligent Systems and Applications in Engineering Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE*, vol. 7, no. 2, pp. 88–91, 2019, doi: 10.1039/b000000x.

[28] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf Sci (N Y)*, vol. 507, pp. 772–794, Jan. 2020, doi: 10.1016/j.ins.2019.06.064.

[29] N. Basuni and Amril Mutoi Siregar, "Comparison of the Accuracy of Drug User Classification Models Using Machine Learning Methods," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 6, pp. 1348–1353, Dec. 2023, doi: 10.29207/resti.v7i6.5401.

[30] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behavioural Processes*, vol. 148, pp. 56–62, Mar. 2018, doi: 10.1016/j.beproc.2018.01.004.

[31] N. S, "Understanding AUC - ROC Curve," 2018.

[32] R. Kannan and V. Vasanthi, "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease," in *SpringerBriefs in Applied Sciences and Technology*, Springer Verlag, 2019, pp. 63–72. doi: 10.1007/978-981-13-0059-2_8.