Journal of Information Systems and Informatics
Vol. 6, No. 3, September 2024 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v6i3.784 Published By DRPM-UBD

Sentiment Analysis of Indonesian Citizens on Electric Vehicle Using FastText and BERT Method

Darryl Rayhan Wijaya[1], Gusti Made Arya Sasmita[2], Wayan Oger Vihikan[3]

[12,3]Department of Information Technology, Udayana University, Bali, Indonesia
Email: [1]darrylrayhanw22@gmail.coml, [2]aryasasmita@unud.ac.id, [3]oger_vihikan@unud.ac.id

Abstract

Electric vehicles have become one of the most important innovations in the automotive industry in recent years. This is not only related to technological developments, but also to its significant impact on the environment and lifestyle of global society. Lot of people do not know about the benefit of using electric vehicles for our environment. The transition from conventional vehicles to electric vehicles can really make our environment healthier and also reducing the pollution. At the same time, debates and feelings about electric vehicles continue to grow around the world. This study aims to understand the dynamics of people's feelings and opinions about electric vehicles through sentiment analysis using the FastText and IndoBERT methods. FastText is an efficient text classification and representation learning method developed by Facebook's AI Research (FAIR) lab. IndoBERT is a pre-trained language model specifically designed for the Indonesian language, leveraging the Bidirectional Encoder Representations from Transformers (BERT) architecture. By analyzing a total of 119,310 data from January 2020 to June 2023, the tweets data were categorized into negative, neutral, and positive classes. Model yielded the highest accuracy of 82.5% using IndoBERT method. The results outcomes positive perceptions of electric vehicles among Indonesian citizen with a percentage of 58%. By carrying out this research, it is hoped that it can produce quality information for producers, the community and the government in developing and advancing public interest in purchasing electric vehicles considering the very positive impact they have on the surrounding environment.

Keywords: Sentiment Analysis, Electric Vehicle, FastText, IndoBERT

1. INTRODUCTION

Electric vehicles are vehicles that use electrical energy as the main power source to drive electric motors. The emergence of electric vehicles is a result of technological advancements in the automotive sector. The sales comparison between electric motorcycles and conventional motorcycles from 2019 to 2022 is vastly different. Sales of electric vehicles during that period only reached 30 thousand units, while conventional vehicles sold 29 million units in the same period [1].

In Indonesia, a country with a large population and rapid economic growth, the implementation of electric vehicles has great potential to contribute to climate change mitigation efforts and sustainable development [2].Electric vehicles have become one of the most important innovations in the automotive industry in recent years. This is not only related to technological developments but also to their significant impact on the environment and the global lifestyle. Electric vehicles have become an attractive alternative to conventional vehicles that run on fossil fuels, because they have several advantages such as being environmentally friendly, energy efficient and lower operating costs[3]. Nowadays, technological development is progressing very rapidly. This is similar to the transition from the use of conventional fuels to electric fuels. However, the adjustment and acceptance of the public towards existing electric motorcycle technology certainly cannot happen in a short time. Public opinion on this matter is often expressed on social media. A popular social media platform for expressing opinions is X (Twitter). According to a We Are Social report, there were 372.9 million X (Twitter) users worldwide as of April 2023. Indonesia ranks sixth with 14.75 million Twitter users. In Indonesia, Twitter is also showing its success with 24 million users in early 2023, covering around 8.7 percent of the country's total population [4].

At the same time, debates and sentiments about electric vehicles continue to evolve worldwide. One way to understand the dynamics of public sentiment and opinion about electric vehicles is through sentiment analysis. Sentiment analysis is a natural language processing technique that allows us to identify, measure, and understand the opinions, feelings, and sentiments expressed by individuals or groups in written text, such as news articles, social media posts, product reviews, or online comments. Sentiment analysis plays an important role in providing consumers with insight into other people's experiences [5].

Sentiment analysis has mostly been done with machine learning and deep learning approaches. Yet, the accuracy of machine learning methods is often insufficient. To enhance precision, deep learning algorithms are now being used [6]. Tweets will be gathered based on related keywords and classified using FastText and BERT.

To gain a deeper and more detailed understanding of public opinion, a classification process was conducted on the X social media platform using deep learning algorithms, specifically the FastText and BERT methods. This approach enables the analysis of tweets gathered through relevant keywords related to the topic of discussion. The classification results will offer more accurate and insightful perspectives on the views and opinions of the Indonesian people regarding electric vehicles.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2. METHODS

This research utilizes several pre-processing stages. The stages used are normalization, case folding, text cleansing, remove symbols, stop word removal, and stemming. Below are the explanations for each stage.

1) Normalization: Converting non-standard words into standard words. The main goal of the normalization process is to standardize all non-standard words so they can be recognized by the model with the same characters [7].
2) Case Folding: Capital letters are converted to lowercase. This is important because capital letters can significantly impact the data analysis process.
3) Text Cleansing: Duplicate data is removed, missing values are filled in, and errors are corrected, which includes addressing spelling mistakes.
4) Remove Symbols: Deleting all symbol, hashtag, and number in the data to make all the data in one form [8].
5) Stop Word Removal: Stop words are common words that do not have significant meaning and are usually ignored during analysis. The purpose of the remove stop words process is to eliminate meaningless words.
6) Stemming: removes affixes from words to be analyzed. The purpose of stemming is to ensure that all words with different affixes are converted into the same word.

The figure below will show and explain the process that will be conducted in this research.

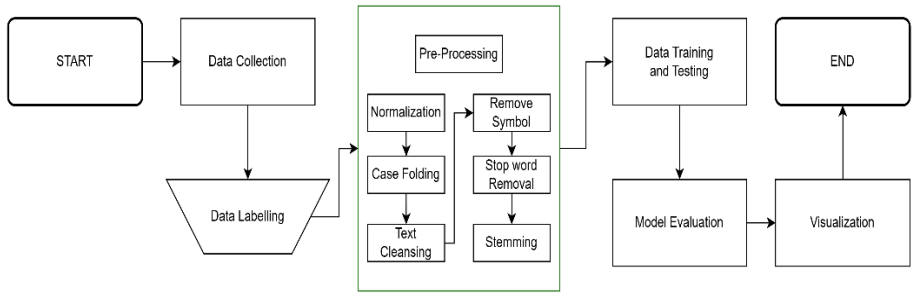

Figure 1. Research Process Flow

Figure 1 outlines the research process, beginning with data collection from social media platform X (Twitter) on topics related to electric vehicles. After gathering the data, 3,000 tweets undergo manual labeling. Both labeled and unlabeled data are then subjected to a six-step pre-processing procedure to clean the data. Only the labeled data is split for further analysis. The model is trained using the cleaned, labeled data, which consists of 3,000 tweets. Next, FastText and BERT models are developed for training and classifying public sentiment. Model evaluation is

performed to identify which method achieves the highest accuracy. Finally, the results are visualized to facilitate easier understanding.

2.1. Data Collection

The data used in this research is from X (Twitter) that obtained by scrapping the tweets using Start by installing the necessary libraries for Tweet Harvest. Next, choose the relevant keywords. When using Tweet Harvest, make sure to input the authentication token from the X (Twitter) account, ensuring that the account is public and not set to private [9]. The keywords of the data scrapping are based around electric vehicle and the date range of the data is from January 2020 until July 2023. The total obtained data is 119,130 data and was saved using CSV format.

2.2. Data Pre-processing

The pre-processing stage aims to clean the data so that it can be ready for use. There are several stages in pre-processing, including normalization, case folding, text cleansing, removing chars, removing stop words, and stemming. The results of each pre-processing stage are shown below.

Table 1. Pre-processing Steps

Steps	Text	Results (English)	Results
Normalization	Jokowi BOHONG lagi. Dulu pernah bilang Esemka sudah dipesen ribuan unit, kini cuap-cuap lagi tentang mobil listrik. #JokowiBohong	Jokowi LIED again. In the past, they said that Esemka had been ordered thousands of units, but now they are talking about electric vehicle #JokowiLied	Jokowi BOHONG lagi. Dulu pernah bilang Esemka sudah dipesan ribuan unit, kini cuap-cuap lagi tentang mobil listrik. #JokowiBohong
Case Folding	Jokowi BOHONG lagi. Dulu pernah bilang Esemka sudah dipesan ribuan unit, kini cuap-cuap lagi tentang mobil listrik. #JokowiBohong	jokowi lied again. in the past, they said that esemka had been ordered thousands of units, but now they are talking about electric vehicle #jokowilied	jokowi bohong lagi. dulu pernah bilang esemka sudah dipesan ribuan unit, kini cuap-cuap lagi tentang mobil listrik. #jokowibohong
Text Cleansing	jokowi bohong lagi. dulu pernah bilang esemka sudah dipesan ribuan unit, kini	jokowi lied again. in the past, they said that esemka had been ordered thousands of units,	jokowi bohong lagi. dulu pernah bilang esemka sudah dipesan ribuan unit,

	cuap-cuap lagi tentang mobil listrik. #jokowibohong	but now they are talking about electric vehicle	kini cuap-cuap lagi tentang mobil listrik.
Remove Symbol	jokowi bohong lagi. dulu pernah bilang esemka sudah dipesan ribuan unit, kini cuap-cuap lagi tentang mobil listrik.	jokowi lied again in the past they said that esemka had been ordered thousands of units but now they are talking about electric vehicle	jokowi bohong lagi dulu pernah bilang esemka sudah dipesan ribuan unit kini cuap cuap lagi tentang mobil listrik
Stop Word Removal	jokowi bohong lagi dulu pernah bilang esemka sudah dipesan ribuan unit kini cuap cuap lagi tentang mobil listrik	jokowi lied and said that esemka ordered thousands of electric car units	jokowi bohong bilang esemka dipesan ribuan unit cuap cuap mobil listrik
Stemming	jokowi bohong bilang esemka dipesan ribuan unit cuap cuap mobil listrik	jokowi lied and said that esemka order thousands of electric car units	jokowi bohong bilang esemka pesan ribu unit cuap cuap mobil listrik

2.3. Data Labelling

Labeled data is tweet data that has gone through the labeling process. In electric vehicle sentiment analysis research, labeling was done manually. Labeled data will be used to test the model. Label data will be labeled manually. The amount of data that will be labeled manually is 3000 data which spread equally to each class. This data will later be used to train the model so that it can produce good accuracy results.

2.4. FastText

FastText is an open source, free, and lightweight library that allows users to learn text representations and text classifiers [10]. FastText is a word embedding method which is a development of word2vec. FastText itself was developed by Facebook. The FastText method will study word representations by considering sub-word information [11]. Each word is represented as a set of n-gram characters. Thus, it can help capture the meaning of shorter words and enable embedding to understand the suffixes and prefixes of words.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

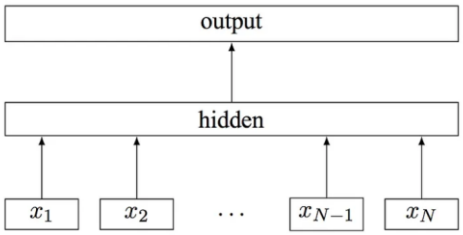

Figure 2. FastText Architecture

Figure 2 is the architecture of FastText. The FastText architecture includes several main layers that facilitate natural language processing. First, the input layer receives words or tokens and converts them into word vectors using the embedding layer. Then, the Bag of Words (BoW) layer combines the word vectors of the entire text, forming a text representation. Hidden layers perform advanced processing of the text representation. What sets FastText apart is its unique approach to building character-based word vectors, allowing the model to address rarely seen or unknown words. Finally, the output layer produces predictions or probability distributions for the output classes in the classification task [12]. With this architecture, FastText can provide strong text representation, especially when dealing with uncommon or unknown words.

2.5. BERT

In recent years, a breakthrough in Natural Language Processing (NLP) known as "Bidirectional Encoder Representations from Transformers" (BERT), has emerged as a powerful and transformative technique, filling the gap in sentiment analysis by going beyond conventional approaches and improving language understanding[13]. Bidirectional Encoder Representations from Transformers (BERT) is a language representation model that was first introduced by Google on October 11 2018. Different from other language models, BERT is designed as a pre-trained model or a model that has been trained bidirectionally from unlabeled text data with aligns the context from both sides of the layer. With this approach, the BERT model can be readjusted by just adding one additional layer [14]. The BERT algorithm has been used by the large company Google since October 21 2021. By using BERT, Google can easily understand natural language text and provide search results that are more relevant and satisfying for its users.

Figure 3 is an image of the BERT architecture. At the Embedding Layer, the input text is converted into an embedding vector representation. Next, in the Transformer Encoder, the embedding of these tokens is then processed through several layers of the Transformer encoder. Each layer consists of a self-attention mechanism and a feed-forward neural network. Self-attention allows each token to pay attention to all other tokens in the sequence, considering context from both directions (before and after). Then the output from the encoder is a vector

representation of each token which is a contextual representation of each token that takes into account the entire sentence. Next is the Classification Layer where the representation of the token is used as a representation of the entire text. The vector is then fed into a classification layer consisting of a Fully-connected layer, GELU (Gaussian Error Linear Unit) activation, and normalization. Then the Softmax function is used to produce probability distributions from various possible classes [15]. In fact, BERT is a neural network-based system. Neural Network is an artificial neural network in machine learning and AI that is used to imitate the working system of the human brain. In NLP that uses Indonesian, it is highly recommended to use IndoBERT compared to BERT.

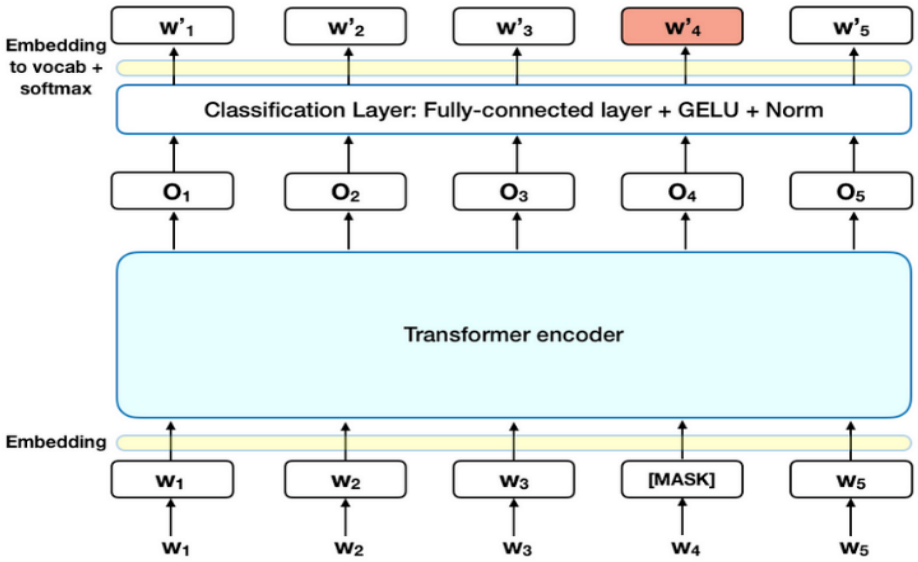

Figure 3. BERT Architecture

3. RESULTS AND DISCUSSION

3.1 Experimental Results

Model testing is a stage that functions to train an algorithm to find out which algorithm has better results. Model testing in electric vehicle sentiment analysis research uses two types of algorithms, namely FastText and IndoBERT. The following is an explanation of the results of each algorithm.

Table 2. FastText Result

Scenario	Accuracy	Precision	Recall	F1-Score
90:10	77.6%	76.1%	77.6%	75.8%
80:20	77.8%	76.8%	77.8%	76.4%

Scenario	Accuracy	Precision	Recall	F1-Score
70:30	76.8%	75.6%	76.8%	75.8%
60:40	76.0%	74.5%	76.0%	74.8%

Table 2 shows the results of testing the FastText model with various scenarios. It can be seen that the highest accuracy was obtained using the 80:20 scenario (80% training data and 20% testing data) with an accuracy of 77.8%. The worst accuracy was obtained using a 60:40 scenario (60% training data and 40% testing data) with an accuracy of 76.0%.

Table 3. IndoBERT Result

Scenario	Accuracy	Precision	Recall	F1-Score
90:10	80.0%	81.8%	80.0%	80.3%
80:20	79.8%	82.0%	79.8%	78.1%
70:30	82.5%	82.4%	82.5%	82.4%
60:40	78.8%	79.0%	78.8%	78.9%

Table 3 shows the results of testing the IndoBERT model with various scenarios. It can be seen that the highest accuracy was obtained using the 70:30 scenario (70% training data and 30% testing data) with an accuracy of 82.5%. The worst accuracy was obtained using a 60:40 scenario (60% training data and 40% testing data) with an accuracy of 78.8%.

The accuracy metric was chosen as a comparison reference because the research carried out can determine which algorithm can produce prediction results that match the actual sentiment class [16]. The research carried out also focuses on dataset scenarios, namely dividing training data and testing data. The main objective of dividing scenarios is to find out which division of testing data and training data can produce the best accuracy [17].

Based on the table above, it can be seen that the best accuracy is owned by the IndoBERT algorithm with a 70:30 scenario which produces an accuracy of 82.5%. The 82.5% accuracy means that the model can predict 82.5% of the data correctly and the remaining 17.5% is predicted incorrectly. Based on these results, electric vehicle sentiment analysis research will use the IndoBERT algorithm with a 70:30 scenario for the next stage, namely the processing stage. This is because the IndoBERT algorithm with the 70:30 scenario can produce the best accuracy compared to other scenarios and other algorithms.

Data visualization is a stage for displaying classification results in a form that is easier to understand [18]. There are several types of data visualization such as line charts, bar charts, pie charts, and others. Visualization is carried out using the dash framework. The following figures are the results of the visualization stage.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

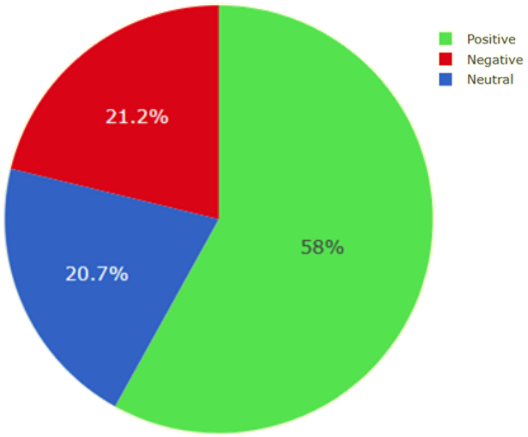

Figure 4. Pie Chart Visualization

Figure 4 is the comparison of all the sentiments in percentage of the Indonesian citizens on the electric vehicle. Most of the people have positive sentiments about electric vehicle which is around 58%. Around 21.2% people have negative sentiments towards electric vehicle. The rest which is around 20.7% were neutral towards the electric vehicle.

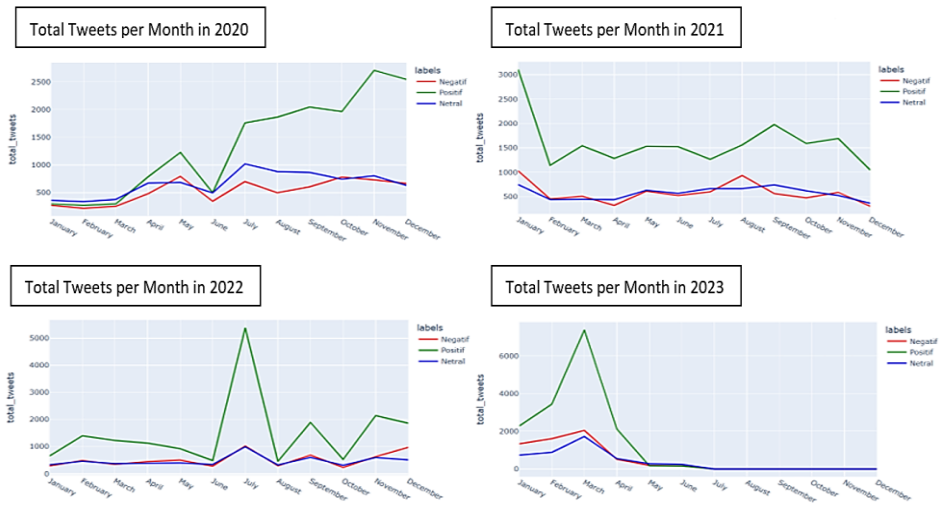

Figure 5. Trend Visualization

Figure 5 is a visualization regarding trends in public sentiment or opinion regarding the presence of electric vehicles, namely from 2020 to 2023. It can be seen in the image above that the trend in public sentiment always changes from year to year and shows a graph that goes up and down.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.2 Discussion

According to Table 2, the IndoBERT algorithm achieved the highest accuracy of 82.50%. This score was achieved using a 70:30 split of the data, with 70% allocated for training and 30% for testing. An accuracy of 82.50% means that the model correctly predicted outcomes for 82.50% of the data, while 17.50% were incorrectly predicted. Based on these findings, the study utilizes the IndoBERT algorithm with a 70:30 data split for the labeling stage because it provided the highest accuracy compared to other scenarios and the FastText algorithm. The superior accuracy of IndoBERT over FastText may be attributed to the data aligning well with the IndoBERT's assumption that features are independent of one another.

Based on Figure 5, in August 2021 the negative trend in society experienced an increase, at which time President Joko Widodo also inaugurated the construction of the first electric vehicle battery factory in Indonesia. This triggers positive and of course negative views from the public. There are many people who tend to agree with the inauguration of the factory because it can advance the electric vehicle industry in Indonesia and can open up new job opportunities. However, people who have negative sentiments think that the government cannot be consistent in advancing the automotive industry in Indonesia, such as what happened with the production of ESEMKA cars in the previous year which was considered a failure. November 2020 was the peak of positive sentiment in 2020 caused by the operation of electric vehicles by the Grab company. Grab claims to have operated more than 5,000 electric vehicles for its fleet in the form of two-wheeled (e-scooters) and four-wheeled vehicles. The increase in sentiment was caused by many people tweeting their opinions when they got grab drivers who used electric vehicles

Figure 6. Word Cloud Visualization

Figure 6 shows the word cloud on each class. The word cloud visualization depicts the most frequent terms used in the context of electric vehicles in two sentiment classes: Positive and Negative. On the left side, labeled "Positive Class," prominent words include "*mobil listrik*" (electric car), "*kendaraan listrik*" (electric vehicle), and "*motor listrik*" (electric motorcycle), indicating a positive sentiment towards various

types of electric vehicles. Other significant terms like "*indonesia*," "*ramah lingkungan*" (environmentally friendly), "*pabrik baterai*" (battery factory), and "*pengisian kendaraan*" (vehicle charging) suggest a positive discourse around the benefits and infrastructure related to electric vehicles in Indonesia.

4. CONCLUSION

This study evaluates the FastText and IndoBERT algorithms for sentiment classification using data from the social media platform X (Twitter). The dataset, collected with TweetHarvest using keywords like "*kendaraan listrik*," "*mobil listrik*," and "*motor listrik*," spans from January 2020 to June 2023 and includes 119,310 tweets. IndoBERT demonstrated the highest accuracy of 82.50% with a 70:30 split, where 70% of the data was used for training and 30% for testing. The sentiment analysis results showed that 58% of the tweets were identified as positive, 21.2% as negative, and 20.7% as neutral. Public sentiment was influenced by various factors including government policies, the adoption of electric vehicles by major companies, exhibitions, and government performance. Peaks in negative sentiment were associated with specific events related to government performance, while positive sentiment spikes correlated with supportive government policies and electric vehicle exhibitions. These findings suggest the need for government efforts to encourage the shift from conventional to electric vehicles for environmental benefits. The research offers valuable insights into public opinion on electric vehicles and demonstrates the comparative effectiveness of FastText and IndoBERT in sentiment analysis.

REFERENCES

[1] D. Ardiyanti, F. Kurniawan, U. Raokter, and R. Wikansari, "Analisis penjualan mobil listrik di Indonesia dalam rentang waktu 2020-2023," *ECOMA J. Econ. Manag.*, vol. 1, no. 3, pp. 114–122, 2023.

[2] R. Merdiansah and A. A. Ridha, "Analisis Sentimen Pengguna X Indonesia Terkait Kendaraan Listrik Menggunakan IndoBERT," *J. Ilmu Komput. Syst. Inf.* (JIKOMSI), vol. 7, pp. 221–228, 2024.

[3] S. Alfarizi and E. Fitriani, "Analisis Sentimen Kendaraan Listrik Menggunakan Algoritma Naive Bayes dengan Seleksi Fitur Information Gain dan Particle Swarm Optimization," vol. 9, no. 1, pp. 19–27, 2023.

[4] A. N. Huzna, I. Nurhayati, A. E. Saputri, and M. Q. Huda, "Analisis Sentimen Terhadap Mobil Listrik Di Indonesia Pada Twitter: Penerapan Naïve Bayes Classifier Untuk Memahami Opini Publik," *Just IT J. Syst. Inf. Technol. Comput.*, vol. 14, no. 2, pp. 87–93, 2024.

[5] G. H. Kusuma, I. Permana, F. N. Salisah, M. Afdal, M. Jazman, and A. Marsal, "Pendekatan Machine Learning: Analisis Sentimen Masyarakat

Terhadap Kendaraan Listrik Pada Sosial Media X," *JUSIFO J. Syst. Inf.*, vol. 9, no. 2, pp. 65–76, 2023, doi: 10.19109/jusifo.v9i2.21354.

[6] D. J. M. Pasaribu, K. Kusrini, and S. Sudarmawan, "Peningkatan Akurasi Klasifikasi Sentimen Ulasan Makanan Amazon dengan Bidirectional LSTM dan Bert Embedding," *Inspir. J. Technol. Inf. Commun.*, vol. 10, no. 1, pp. 9–20, 2020, doi: 10.35585/inspir.v10i1.2568.

[7] N. M. T. O. Adriana, I. M. A. D. Suarjaya, and D. P. Githa, "Analisis Sentimen Publik Terhadap Aksi Demonstrasi di Indonesia Menggunakan Support Vector Machine Dan Random Forest," *Decode J. Educ. Technol. Inf.*, vol. 3, no. 2, pp. 257–267, 2023, doi: 10.51454/decode.v3i2.187.

[8] A. Srivastava, V. Singh, and G. S. Drall, "Sentiment analysis of twitter data: A hybrid approach," *Int. J. Healthc. Inf. Syst. Inform.*, vol. 14, no. 2, pp. 1–16, 2019, doi: 10.4018/IJHISI.2019040101.

[9] F. M. J. M. Shamrat et al., "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 1, pp. 463–470, 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.

[10] M. A. A. Islamy and P. P. Adikara, "Analisis Sentimen IMDB Movie Reviews menggunakan Metode Long Short-Term Memory dan FastText," vol. 6, no. 9, pp. 4106–4115, 2022.

[11] R. P. Hastuti, V. Riona, and M. Hardiyanti, "Content Retrieval dengan Fasttext Word Embedding pada Learning Management System Olimpiade," *J. Internet Softw. Eng.* (JISE), vol. 4, no. 1, 2023.

[12] Y. Fang, Y. Liu, C. Huang, and L. Liu, "Fastembed: Predicting vulnerability exploitation possibility based on ensemble machine learning algorithm," *PLoS One*, vol. 15, no. 2, pp. 1–28, 2020, doi: 10.1371/journal.pone.0228439.

[13] M. S. Sayeed, V. Mohan, and K. S. Muthu, "BERT: A Review of Applications in Sentiment Analysis," *HighTech Innov. J.*, vol. 4, no. 2, pp. 453–462, 2023, doi: 10.28991/HIJ-2023-04-02-015.

[14] J. D. M. W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Jun. 2019, vol. 1, p. 2.

[15] M. F. A. Islamy, D. Richasdy, and M. A. Bijaksana, "BERT Implementation on News Sentiment Analysis and Analysis Benefits on Branding," vol. 6, pp. 2064–2073, 2022, doi: 10.30865/mib.v6i4.4579.

[16] R. Kurniawan, F. Lestari, A. S. Batubara, M. Z. A. Nazri, K. Rajab, and R. Munir, "Indonesian Lexicon-Based Sentiment Analysis of Online Religious Lectures Review," in *2021 Int. Congr. Adv. Technol. Eng.* (ICOTEN), Jul. 2021, pp. 1–5, IEEE.

[17] T. K. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: A review," in *2017 Int. Conf. Invent. Commun. Comput. Technol.* (ICICCT), Mar. 2017, pp. 298–301, IEEE.

[18] W. Irmayani, "Visualisasi Data Pada Data Mining Menggunakan Metode Klasifikasi Naive Bayes," *J. Khatulistiwa Inform.*, vol. 9, no. 1, 2021.