BIS INFORMATION TECHNOLOGY AND COMPUTER SCIENCE

Volume 1 2024 | DOI: 10.31603/bistyc.140
5th Borobudur International Symposium on Science and Technology (BIS-TYC) 2023

UNIMMA PRESS
UNIVERSITAS MUHAMMADIYAH MAGELANG

Model machine learning for sentiment analysis of the presence of electric vehicle in Indonesia

A M Siregar[1*], S Faisal[1], A Fauzi[1], J Indra[1], A F N Masruriyah[1] and A R Pratama[1]

[1] Department of Informatics Engineering, Faculty of Computer Science, University of Buana Perjuangan, Karawang, Indonesia

[*]Corresponding author email: amrilmutoi@ubpkarawang.ac.id

Abstract

Sentiment analysis, also known as opinion analysis or social sentiment analysis, is a well-established field of study. Within the automotive industry, great attention is being paid to the presence of electric cars as a viable solution to the pressing issue of greenhouse gas emissions. In order to gauge the level of acceptance and adoption of this technology, it is crucial to analyze the sentiments and opinions expressed by individuals towards electric cars Various approaches can be employed for sentiment analysis, including rule-based techniques, statistical methods, and machine learning algorithms. The objective of this research endeavor is to conduct sentiment analysis on online publications and social media discussions pertaining to electric cars. Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) are the specific methods employed in this study. The effectiveness of these methods is evaluated using accuracy measurements and Receiver Operating Characteristic (ROC) analysis. The accuracy outcomes attained by LR were 78.02%, SVM achieved 71.92%, and RF exhibited 82.35%. By virtue of the examination outcomes of multiple techniques utilized, there is an optimistic expectation that this can serve as the initial stride towards constructing sentiment applications for the existence of electric cars in the Indonesian context.

Keywords

Machine learning, Sentiment analysis, Electric vehicle

Published:
October 20, 2024

Selection and Peer-review under the responsibility of the 5th BIS-STE 2023 Committee

Introduction

According to data provided by IQAir, the pollution level in Jakarta reached 163 AQI US at 09.00 WIB, placing it as the fifth most polluted city in the world. This level of pollution indicates that the air in Jakarta is unsuitable for the well-being of the community, therefore it is advisable to wear a mask when engaging in outdoor activities. Kolkata, India currently holds the unenviable position of having the most severe air pollution in the world, with a US AQI of 247. It is closely followed by Dakha, Bangladesh with a US AQI of 208, Delhi, India with a US AQI of 177, Lahore, Pakistan with a US AQI of 163, Jakarta, Indonesia with a US AQI of 163, Ulaanbaatar, Mongolia with a US AQI of 162,

and Kuwait City, Kuwait with a US AQI of 160. Meanwhile, the con- centration level of PM2.5 stands at 79.5 µg/m³, which is equivalent to 15.9 times the annual air quality guideline value set by the World Health Organization (WHO). The morning temperature in Jakarta registered at 29 degrees Celsius, accompanied by a humidity level of 79%. Additionally, wind motion was observed at a rate of 5.5 km/h and atmospheric pressure stood at 1012 mbar. It is noteworthy to mention that Jakarta currently holds the 5th position in the global air quality ranking, with a red indicator signifying an unhealthy condition.

In order to diminish the extent of air contamination in Indonesia, with particular emphasis on the city of Jakarta, one of the endeavors employed is the utilization of electric vehicles, which has garnered global attention as a more ecologically sound alternative within the transportation domain [1][2]. In an endeavor to abate emissions and enhance the sustainability of the transportation sector, the Indonesian government has implemented policies and incentives designed to stimulate the populace to utilize electric vehicles. In accordance with the Ministry of Industry (MOI), as of September 2022, the quantity of electric vehicles (excluding hybrid or other variations) in Indonesia has sur- passed the threshold of 25,000 units.

In the realm of electric vehicle implementation, there exists a multitude of viewpoints within the community. The opinions and sentiments surrounding the utilization of electric vehicles showcase a remarkable diversity [3][4][5]. Certain individuals ardently advocate for electric vehicles, perceiving them as a formidable solution to mitigate pollution and decrease reliance on fossil fuels. Advocates contend that electric vehicles possess the potential to curtail greenhouse gas emissions, enhance air quality, and di- minish dependency on petroleum [6]. Conversely, there are skeptics who express reservations regarding the adequacy of battery charging infrastructure, limited range, and prolonged charging duration in comparison to conventional refueling [7].

Sentiments regarding the utilization of electric vehicles are frequently articulated through online networking platforms [8]. Online networking users possess the capability to directly disseminate their perspectives via posts, comments, or by employing pertinent hashtags, and online networking serves as a noteworthy medium for individuals to exchange their viewpoints, encounters, and sentiments on an assortment of subjects, encompassing the adoption of electric vehicles [9][10]. Consequently, sentiment analysis predicated on online networking can prove to be an exceedingly advantageous instrument for comprehending the public's perspectives and attitudes towards the employment of electric vehicles in Indonesia.

Previous research has been carried out within the realm of sentiment analysis on social media pertaining to the adoption of electric vehicles in different nations. To illustrate, a sentiment analysis study conducted in China by [11] employed social media as a basis, establishing that a majority of sentiments associated with electric vehicles were positive. This was attributed to factors such as environmental cleanliness, energy efficiency, and advanced technology. Conversely, a study conducted in the United

States by [12] discovered that the public sentiment towards electric vehicles exhibits considerable variability, encompassing both positive aspects, such as sustainability, and negative aspects, such as cost and the availability of charging infrastructure.

The aim of this study is to carry out sentiment analysis of social media discussions, concentrating on the utilization of electric vehicles in Indonesia and to amass data from the social media platform Twitter. This investigation proposes a machine learning technique involving three algorithms, specifically support vector machine (SVM), Logistic regression (LR), and Random Forest (RF), with the intention of obtaining the most optimal model that will be implemented into a sentiment analysis application designed explicitly for the existence of electric vehicles. The categories of public sentiment include positive, negative, and neutral. The neutral sentiment exhibits lesser concern regarding the presence of electric vehicles, aiming to decrease carbon emissions.

Methods

In this study, a series of stages were implemented within the process. Commencing with the initial stage, data collection, or rather, "data crawling", was conducted. Subsequently, the subsequent stage entailed data preprocessing. This was later succeeded by the data labeling stage. The subsequent stage involved the modeling process, which was proceeded by model evaluation. A comprehensive depiction of the research flow is visually presented in Figure 1.

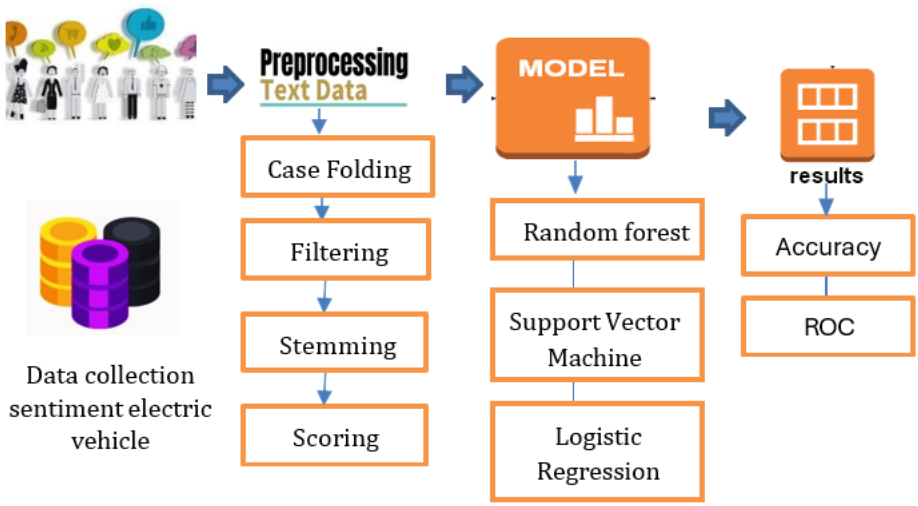


Figure 1. Flow of the proposed research

Data collection

In this stage, information is obtained from the Twitter social media platform. The process of acquiring data will involve the application of web scraping techniques to extract information from Twitter [13]. The acquired information comprises the general sentiment of the public with regards to the utilization of electric vehicles in Indonesia. The designated keyword employed to gather data is "electric vehicles". The process of gathering data was executed through the utilization of Google Colab with the intention

of facilitating the retrieval of information. The amount of data retrieved subsequent to the cleanup process amounted to 4579, which was then reduced to a total of 4502.

Pre-processing data

After the successful attainment of the data, the subsequent procedure entails the pre-processing of the data in order to derive significant attributes from each individual tweet. The purpose of pre-processing lies in the elimination of extraneous noise or inconsequential details [14]. Various steps pertaining to pre-processing shall be executed as outlined below.

Case folding

By implementing the process of case folding, all alphabetical characters within the text of a tweet will undergo a transformation into lowercase letters. This conversion occurs while preserving the original meaning and sentiment encapsulated within the text [15], [16]. This particular technique serves to diminish superfluous discrepancies stemming from different fonts found in the tweet data. Consequently, it facilitates a more consistent and precise analysis of sentiment, thereby assisting subsequent stages of analysis, including tokenization.

Filtering

Filtering constitutes a crucial initial stage in the process of selecting the attributes to be employed. Within the realm of filtering, attributes that lack informative value or fail to make a noteworthy impact are eliminated. This process is undertaken to concentrate attention on the most pertinent attributes and enhance comprehension of public sentiment surrounding the adoption of electric vehicles [17][6].

Stemming

Stemming is employed to transform words into their fundamental or foundational state. For instance, words like "read", "recite", and "read-read" (read repeatedly) will be transformed into the foundational state "read" [18]. Through the application of stemming, researchers are able to simplify the word portrayal of the word portrayal in the accumulated tweet texts.

Scoring

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic that reflects how important a word is to a document in a collection or corpus. It is commonly used in information retrieval and text mining. Scoring TF-IDF involves calculating two main components: Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency is a measure of how often a term appears in a document. It is calculated as the number of times a term appears in a document divided by the total number of terms in that document. TF(t,d) = Number of times term t appears in document d Total number of terms in document dTF (t,d) = Total number of terms in document d Number of times term t appears in document d. The higher the TF-IDF score, the more important the term is in the document or corpus [19][13].

Machine learning model

At this phase, the processed information will be employed to train the model utilizing the Support Vector Machine (SVM) approach with hyperplane. SVM constructs a hyperplane (e.g., line or surface) to segregate the information among distinct categories. The aim is to discover the optimal hyperplane that exhibits the maximum margin be- tween the different categories in the information [20]. Logistic regression is a statistical approach utilized for binary classification, forecasting the likelihood of an occurrence fitting into a specific category. Despite its given title, logistic regression is employed for classification, not regression. It proves to be particularly advantageous when the reliant variable is categorical and comprises two categories (binary classification), such as spam or non-spam, pass or fail [14]. Random Forest is an ensemble learning technique that functions by constructing numerous decision trees during the training phase and producing a class that represents the mode of the classes (for classification) or the average prediction (for regression) of each tree. Random Forest is an influential and adaptable algorithm in the field of machine learning, possessing various noteworthy features [21]. This procedure will facilitate the examination of individuals' viewpoints and sentiments regarding the adoption of electric vehicles in Indonesia.

Evaluation model

The process of evaluating a model entails the utilization of pertinent evaluation measures that are applicable to the assessment of model performance [22]. As an ex- ample, evaluation measures that are commonly employed encompass accuracy, precision, recall, and F1-score. Accuracy gauges the model's ability to accurately classify sentiments. Precision gauges the veracity of positive results classified by the model, while recall gauges the model's capacity to identify and capture all correct positive results. Receiver Operating Characteristic (ROC) curves are visual representations of the performance of binary classification models at various classification thresholds. These curves illustrate the compromise between the rate of true positives (sensitivity or recall) and the rate of false positives (1 - specificity) for different threshold values [23].

Results and Discussion

Depending on the context and purpose of the research, sentiment analysis can consider numerous factors. The process typically entails the collection and analysis of textual data to ascertain the opinion or sentiment pertaining to a particular subject or topic. Potential research outcomes stemming from sentiment analysis encompass the determination of whether the prevailing sentiment towards a given topic is positive, negative, or neutral. Additionally, the identification of specific patterns or trends in sentiment over time or among specific groups may also emerge. Another possibility is the classification of sentiment into distinct categories, including positive, negative, or neutral. In cases where the research involves the development of sentiment models,

such as machine learning, the efficacy and accuracy of these models in predicting sentiment are evaluated.

Results

The dataset outcomes attained in this investigation were acquired from the social media platform, Twitter. As depicted in Figure 2, the visualization portrays that a majority of the sentiments expressed were neutral, accounting for 57.1%, while positive sentiments constituted 16.1% and negative sentiments made up 26.8%. This data signifies that the neutral sentiment prevails and implies that the citizens of Indonesia either refrain from expressing their opinion or do not prioritize the presence of electric vehicles as a means to mitigate carbon emissions.

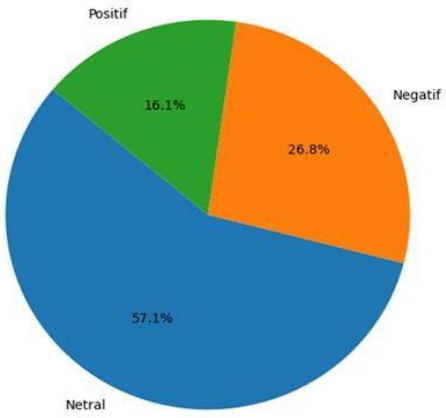


Figure 2. Proportional number of datasets used

Through the illustration depicted in Figure 3, it becomes evident that the frequency of occurrence of the most popular words is concentrated within the top 10. These words include battery, *ngecas*, *batrenya*, token, charger, watt, *lembung*, powerbank, *nge*charge, and KWH. The visual representation exhibits that the top 10 words possessing the highest level of centrality are *ngecas*, battery, *batrenya*, token, charger, *lembung*, powerbank, watt, *ngechasnya*, and KWH. In Figure (not specified), the words maintain a nearly identical sequence.

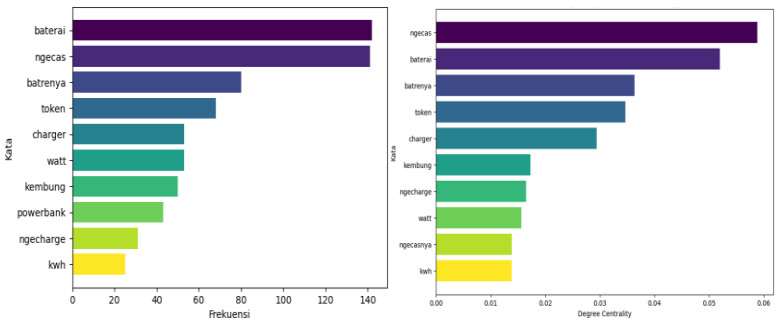
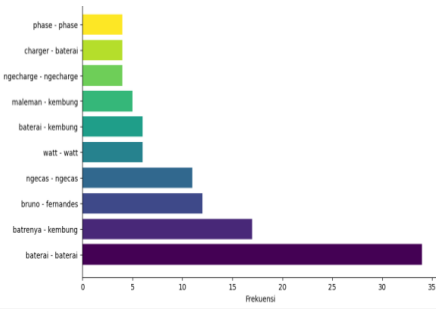
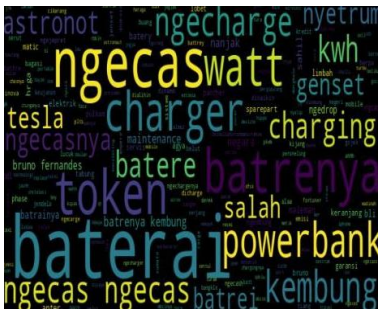


Figure 3. Proportional number of datasets used

The visualization depicted in Figure 4 reveals that the top 10 most popular words include phase-phase, charger-battery, *ngecharge-ngecharge*, *maleman*-bloated, watt-watt, ngecas-ngecas, and others. Figure 4b visually represents the largest word in terms of frequency in sentiment analysis, with battery, charge, token, charges, and others being the predominant words.



a. Top 10 words pairs



b. Visualization of wordcloud

Figure 4. Number of word pairs in sentiment and wordcloud

In Figure 5, the outcomes of the assessment of machine learning algorithms' performance in classifying sentiment regarding the existence of Indonesian electric vehicles using the ROC method are presented. The sentiment categories encompass three options: positive, negative, and neutral. The Random Forest algorithm succeeded in attaining the most optimal model, attaining a level of 80% accuracy. On the other hand, the SVM algorithm achieved a level of accuracy of 62%, while the logistic regression algorithm reached an accuracy level of 77%.

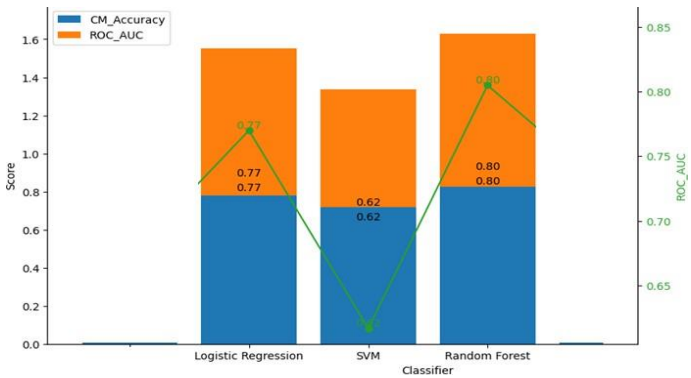


Figure 5. Evaluating the performance of machine learning with ROC

In the Table 1, we present the outcomes of the performance assessment regarding the utilization of a machine learning algorithm for categorizing sentiments associated with electric vehicles. The sentiments are divided into three distinct categories, namely positive, negative, and neutral. Amongst the various algorithms employed, the Random Forest algorithm demonstrated the highest level of efficacy, yielding a model with an accuracy rate of 82.35%. Similarly, the SVM algorithm achieved a commendable accuracy rate of 71.92%, while the logistic regression algorithm attained an accuracy rate of 78.02%.

Table 1. performance of machine learning

Algorithms	Accuracy
Logistic Regression	78.02%
Support Vector Machine	71.92%
Random Forest	82.35%

Discussion

After completing all the necessary steps of this research, which include data collection, model evaluation, and identification of the most advantageous model among the

employed algorithms, it has become apparent that Indonesian individuals exhibit limited interest in engaging in discussions pertaining to social media platforms. This is substantiated by the observation that the topic of electric cars fails to elicit comments. Specifically, in the context of the social media conversations that were gathered, it was found that more than 57% of individuals did not provide any response to the introduction of electric cars in Indonesia. It is noteworthy that the government has implemented a range of policies aimed at promoting the adoption of electric vehicles, such as tax exemptions, the expansion of charging infrastructure, and the simplification of vehicle ownership certification processes.

The most advantageous model, as determined by the random forest algorithm, demonstrates an accuracy rate of 82.35%, surpassing the performance of both the SVM and Logistic regression algorithms. This optimal model could be further utilized in the development of applications that gauge public sentiment towards the presence of electric vehicles. Consequently, the government would be empowered to devise policies that effectively stimulate public concern and interest in this mode of transportation.

Conclusion

Based on the research findings, it is possible to draw various conclusions. One such conclusion is that Indonesians display a lack of concern when it comes to expressing their opinions about electric vehicles through social media. This is evident from the fact that more than 50% of conversations observed were of a neutral nature. Moreover, the accuracy results achieved by LR were 78.02%, SVM attained 71.92%, and RF exhibited 82.35%. These outcomes highlight the effectiveness of the multiple techniques employed in the examination. Consequently, there exists a positive expectation that these findings can serve as an initial step towards developing sentiment applications for the presence of electric cars in the Indonesian context.

References

[1] K. Kolbe, "Mitigating urban heat island effect and carbon dioxide emissions through different mobility concepts: Comparison of conventional vehicles with electric vehicles, hydrogen vehicles and public transportation," Transp. Policy, vol. 80, pp. 1–11, Aug. 2019, doi: 10.1016/j.tranpol.2019.05.007.

[2] X. Hu, R. Zhou, S. Wang, L. Gao, and Z. Zhu, "Consumers' value perception and intention to purchase electric vehicles: A benefit-risk analysis," Res. Transp. Bus. Manag., vol. 49, p. 101004, Aug. 2023, doi: 10.1016/j.rtbm.2023.101004.

[3] M. I. Alhari, O. N. Pratiwi, and M. Lubis, "Sentiment Analysis of The Public Perspective Electric Cars in Indonesia Using Support Vector Machine Algorithm," in 2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA), Nov. 2022, pp. 155–160, doi: 10.1109/ICSINTESA56431.2022.10041604.

[4] N. Ashari, M. Z. Mifta Al Firdaus, I. Budi, A. B. Santoso, and P. Kresna Putra, "Analyzing Public Opinion on Electrical Vehicles in Indonesia Using Sentiment Analysis and Topic Modeling," in 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), Feb. 2023, pp. 461–465, doi: 10.1109/ICCoSITE57641.2023.10127834.

[5] M. Wang, H. You, H. Ma, X. Sun, and Z. Wang, "Sentiment Analysis of Online New Energy Vehicle Reviews," Appl. Sci., vol. 13, no. 14, p. 8176, Jul. 2023, doi: 10.3390/app13148176.

[6] M. Lu, X. Zhang, J. Ji, X. Xu, and Y. Zhang, "Research progress on power battery cooling technology for electric vehicles," J. Energy Storage, vol. 27, p. 101155, Feb. 2020, doi: 10.1016/j.est.2019.101155.
[7] T. Capuder, D. Miloš Sprčić, D. Zoričić, and H. Pandžić, "Review of challenges and assessment of electric vehicles integration policy goals: Integrated risk analysis approach," Int. J. Electr. Power Energy Syst., vol. 119, p. 105894, Jul. 2020, doi: 10.1016/j.ijepes.2020.105894.
[8] R. Jena, "An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach," Ind. Mark. Manag., vol. 90, pp. 605–616, Oct. 2020, doi: 10.1016/j.indmarman.2019.12.012.
[9] L. M. Austmann and S. A. Vigne, "Does environmental awareness fuel the electric vehicle market? A Twitter keyword analysis," Energy Econ., vol. 101, p. 105337, Sep. 2021, doi: 10.1016/j.eneco.2021.105337.
[10] S.-C. Ma, Y. Fan, J.-F. Guo, J.-H. Xu, and J. Zhu, "Analysing online behaviour to determine Chinese consumers' preferences for electric vehicles," J. Clean. Prod., vol. 229, pp. 244–255, Aug. 2019, doi: 10.1016/j.jclepro.2019.04.374.
[11] Q. Qin, Z. Zhou, J. Zhou, Z. Huang, X. Zeng, and B. Fan, "Sentiment and attention of the Chinese public toward electric vehicles: A big data analytics approach," Eng. Appl. Artif. Intell., vol. 127, p. 107216, Jan. 2024, doi: 10.1016/j.engappai.2023.107216.
[12] S. Hardman, R. Berliner, and G. Tal, "Who will be the early adopters of automated vehicles? Insights from a survey of electric vehicle owners in the United States," Transp. Res. Part D Transp. Environ., vol. 71, pp. 248–264, Jun. 2019, doi: 10.1016/j.trd.2018.12.001.
[13] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," Expert Syst. Appl., vol. 66, pp. 245– 260, Dec. 2016, doi: 10.1016/j.eswa.2016.09.009.
[14] T. H. Jaya Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," Procedia Comput. Sci., vol. 197, pp. 660–667, 2022, doi: 10.1016/j.procs.2021.12.187.
[15] W. Bourequat and H. Mourad, "Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine," Int. J. Adv. Data Inf. Syst., vol. 2, no. 1, pp. 36–44, Apr. 2021, doi: 10.25008/ijadis.v2i1.1216.
[16] S. Sathyan, J. J. Peedikayil, R. P. V, and S. R. Salkuti, "Two-Layered Machine Learning Approach for Sentiment Analysis of tweets related to Electric Vehicles," in 2023 International Conference on Innovations in Engineering and Technology (ICIET), Jul. 2023, pp. 1–6, doi: 10.1109/ICIET57285.2023.10220717.
[17] V. Breschi, M. Tanelli, C. Ravazzi, S. Strada, and F. Dabbene, "Social network analysis of electric vehicles adoption: a data-based approach," in 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Sep. 2020, pp. 1–4, doi: 10.1109/ICHMS49158.2020.9209373.
[18] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre- trained language model and model averaging," J. Big Data, vol. 8, no. 1, p. 68, Dec. 2021, doi: 10.1186/s40537-021-00459-1.
[19] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," Expert Syst. Appl., vol. 38, no. 3, pp. 2758–2765, Mar. 2011, doi: 10.1016/j.eswa.2010.08.066.
[20] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.
[21] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 6, pp. 2733–2742, Jun. 2022, doi: 10.1016/j.jksuci.2022.03.012.
[22] A. M. Siregar, Y. A. Purwanto, S. H. Wijaya, and N. Nahrowi, "Two-stages of segmentation to improve accuracy of deep learning models based on dairy cow morphology," Int. J. Electr. Comput. Eng., vol. 13, no. 2, p. 2093, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2093-2100.
[23] J. T. Wixted and L. Mickes, "Evaluating eyewitness identification procedures: ROC analysis and its misconceptions.," J. Appl. Res. Mem. Cogn., vol. 4, no. 4, pp. 318–323, Dec. 2015, doi: 10.1016/j.jarmac.2015.08.009