BEEI-03April2024-FinalPaper

by Kristiawan Nugroho

Submission date: 03-Apr-2024 10:12AM (UTC+0700)
Submission ID: 2338520433
File name: Paper_kristiawan-BEEI-FinalPaper03April2024.docx (237.33K)
Word count: 4757
Character count: 26836

Bulletin of Electrical Engineering and Informatics
Vol. 99, No. 1, Month 2099, pp. 1~1x
ISSN: 2302-9285, DOI: 10.11591/eei.v99i1.paperID

Enhanced Multi-lingual Twitter Sentiment Analysis Using Hyperparameter Tuning K-Nearest Neighbors

Kristiawan Nugroho[1], Edy Winarno[1], De Rosal Ignatius Moses Setiadi[2], Omar Farooq[3]
[1]Faculty of Information and Industrial Technology, Universitas Stikubank, Semarang, Indonesia
[2]Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[3]Faculty of Electronics Engineering, Z. H. College of Engg. and Technology, A. M. U. Aligarh, India

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Social Media
Twitter
Sentiment Analysis
KNN
Hyperparameter Tuning

ABSTRACT

Social media is a medium that is often used by someone to express themselves. These various problems on social media have encouraged research in sentiment analysis to become one of the most popular research fields. Various methods are used in sentiment analysis research, ranging from classic machine learning to deep learning. Researchers nowadays often use deep learning methods in sentiment analysis research because they have advantages in processing large amounts of data and providing high accuracy. However, deep learning also has limitations on the longer computational side due to the complexity of its network architecture. KNN is a robust machine learning method but does not yet provide high-accuracy results in multi-lingual sentiment analysis research, so a hyperparameter tuning KNN approach is proposed. The results showed that using the proposed method, the accuracy level improved to 98.37%, and the classification error improved to 1.63%. The model performed better than other machine learning and even deep learning methods. The results of this study indicate that KNN using hyperparameter tuning is a method that contributes to the sentiment analysis classification model using the Twitter dataset.

Corresponding Author:

Kristiawan Nugroho
Faculty of Information and Industrial Technology
Universitas Stikubank
Jl. Tri Lomba Juang, Kota Semarang, Indonesia
Email: kristiawan@edu.unisbank.ac.id

1. INTRODUCTION

Sentiment analysis is an interesting research topic because it relates to the expression of social conditions in society, including economic, political, cultural, and technological developments. Sentiment analysis is a data analysis technique used to identify and extract sentiments or opinions from text, such as product reviews, tweets, or blog posts. Sentiment analysis is one of society's trending and interesting topics [1]–[3], so some researchers are still interested in exploring various sides to find the latest novelty.

Sentiment analysis, which is part of natural language processing(NLP), has become a popular research topic for several reasons, namely the increasing use of social media, so that more and more people are using social media to interact with friends, family and products[4]. In the economic and business fields, sentiment analysis can help companies understand user sentiment towards their brands and products. In the field of e-commerce, Singh[5], Gondhi[6], and Sinnasamy[7] conducted product review research using various methods in computer science that are able to identify customer tastes and tendencies in using various kinds of products.

In addition, research in the field of sentiment analysis is used to help companies find new business opportunities or existing opportunities that have not been exploited so that companies can diversify products that will be beneficial for them in marketing their new products. Sentiment analysis can also be used to improve customer experience. Companies that understand customer sentiment can adapt their products or services to meet customer needs and enhance the customer experience.

Machine Learning (ML) is part of artificial intelligence technology whose job is to provide learning on computer systems to produce models that can be used for classification, predictions and data clustering[8]–[10]. ML is an approach often used by researchers in analyzing sentiment on social media. Various classical methods in machine learning have been used and have proven successful in helping classify sentiment analysts for product reviews[11], [12], service reviews[13],[14], reviews of places[15],[16] and hoax news[17]. Machine learning methods often used include Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbor.

However, the various methods used have not been able to provide maximum accuracy in classifying sentiment analysis. The reason is that sentiment analysis requires large and varied data for good model training. If the model is only trained on a small amount of data or limited data, then the accuracy of the model will not be maximized. In addition, some models will rely heavily on training data, which can affect the model's accuracy. If the training data does not represent the actual data or is not varied enough, then the model will not be able to generalize well to data that has never been seen before.

As a solution to this problem, to achieve maximum accuracy in sentiment analysis with machine learning, a lot of high-quality training data is needed, as well as a model that can adapt well to data that changes over time and to subjective and ambiguous language characteristics. The Deep Learning approach is widely used to process large training data, which is expected to increase accuracy optimally. However, various methods in deep learning require expensive computational processes, especially in terms of dataset training time and modern infrastructure for managing these datasets.

K-Nearest Neighbor (KNN) is a method in Machine Learning that is robustly used in various types of research. KNN is one of the simplest and most popular machine learning algorithms for classification and regression. Other advantages of the KNN algorithm include being simple and easy to implement because KNN does not require a complex training process like other machine learning algorithms. In addition, KNN is effective for high-dimensional data: K-NN is more effective than other machine learning algorithms on high-dimensional data.

KNN has been used in various studies combined with other approaches aimed at producing more optimal model accuracy performance, such as research conducted by Seikhi[18], who conducted research using KNN to measure the accuracy of several datasets, including Wine with an accuracy reaching 94.33% with K=7 to 96.22% with K=3. on the other hand Kadry[19] succeeded in combining KNN with PSO (Particle Swarm Optimization) which achieved an accuracy rate of 88.59%. Enriko[20]also uses KNN to predict heart disease, which produces an accuracy of 81.9%. Various studies that have been carried out using KNN try to hybridize several approaches to achieve a maximum level of accuracy, one of which is the hyperparameter tuning method.

Hyperparameter tuning, also known as hyperparameter optimization, refers to selecting the best values for the hyperparameters of a machine-learning model. It systematically explores different combinations of hyperparameter values to find the optimal configuration that yields the best performance for a specific task or problem. The hyperparameter tuning approach is also applied to the KNN algorithm, including in research conducted by Wazirali[21] in network security on the topic of intrusion detection. The hyperparameter cross-validation model produces a model accuracy rate of 98.87%. In the health sector, Ambesange[22] used the KNN with the hyperparameter technique in detecting liver disease using the Indian Liver Patients dataset, this approach has succeeded in producing an accuracy rate of up to 93%.

In the research that has been done, KNN produces less than optimal accuracy performance with an accuracy below 90%. This study proposes an approach to improve the KNN method by using the hyperparameter tuning technique on the Twitter Sentiment Multi-Lingual dataset so that it will produce a high level of accuracy. Setting hyperparameters in the k-NN algorithm requires experimentation and a good understanding of data characteristics. Through iteration and validation, the optimal hyperparameter combination can be found for the goal to be achieved with the k-NN algorithm.

In this work the writing of the paper is divided into several parts, namely the introduction which contains the background to the problem, the state of the art and the proposed method. The next section is Method which discusses the methods used in this research and the proposed method. Next is the Research Results and Discussion which describe the results of the research that has been carried out and comparisons with several similar studies. The last section is the conclusion which discusses the conclusions resulting from the research and future research that will be carried out.

2. METHOD

Research methods are used in science to obtain data, gather information, analyze phenomena, and test hypotheses. Using the Machine Learning approach as a research method, researchers can leverage powerful computing capabilities to process data, identify complex patterns, predict future behavior, and generate valuable insights in various research fields. Research on multilingual Twitter sentiment analysis was carried out using the research stages, which can be seen in Figure 1 as follows:

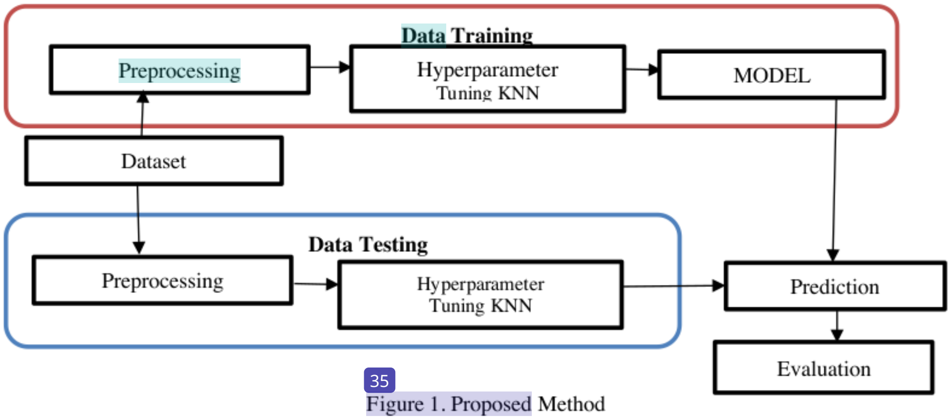


Figure 1. Proposed Method

2.1. Data Collection

This stage is the first step in research on machine learning. Dataset collection gathers relevant and representative data from various sources for analysis or research purposes. According to Taherdoost[23], data collection is useful for gaining insight into research topics. Datasets can consist of numeric, text, or image data and can be collected through various means, such as surveys, interviews, observations, or experiments. In this study, the Twitter Sentiment Analysis dataset was taken from the site https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analisis.

The Twitter Sentiment Analysis dataset consists of 2 files, namely twitter_training.csv, which is the dataset used to conduct model training which consists of 74,139 rows of data, and twitter_validation.csv, which is the dataset to validate the resulting model, which consists of 1757 data. This dataset consists of five features: row no,att1,att2,att3, and att4 with att2, a class containing positive, negative, and neutral sentiments.

2.2. Preprocessing

Data preprocessing is a crucial step that must be carried out in research[24]. Data preprocessing is a series of steps performed on the dataset before analysis or modeling. The purpose of dataset preprocessing is to clean, modify, and prepare the dataset so that it is ready for analysis or modeling. This study uses some preprocessing activities. Some of the preprocessing stages carried out are:

2.2.1 Remove Duplicates

This feature is used to delete the same data. Initially, there were 74,139 data, but after using the remove duplicates feature, there are currently only 71,792 data ready to be processed. The data before and after the preprocessing can be seen in Table 1.

Table 1. Data Preprocessing - Remove Duplicates

Row no	Att1	Att2	Att3	Att4
1	2401	Borderlands	Positive	Im getting on…
2	2401	Borderlands	Positive	Iam coming t…

Before Preprocessing (74139 examples, 1 special attribute,3 regular attributes)

Row no	Att1	Att2	Att3	Att4
71791	9200	Nvidia	Positive	Just realized..
71792	9200	Nvidia	Positive	Just like the..

After Preprocessing (71792 examples, 1 special attribute,3 regular attributes)

Table 1 shows the preprocessing process to remove duplicate data so that the data to be processed is cleaner. In the context of data processing, the "remove duplicate" step functions to remove identical or duplicate entries or rows in the dataset. The main goal is to clean the data and prevent unnecessary redundancy, which can affect the analysis and models built from the dataset.

2.2.2 Replace Missing Values

The next preprocessing stage is to replace empty feature values with certain values. This stage is an important step that must be taken so that all features are filled with all the required data. The following is an example of data that will be processed by looking for empty values and replacing them.

Table 2. Data Preprocessing – Replace Missing Values

Row no	Att1	Att2	Att3	Att4
57	2410	Negative	Borderlands	why does pra…
58	2411	Neutral	Borderlands	….[
59	2411	Neutral	Borderlands	.?
60	2411	Neutral	Borderlands	
61	2411	Neutral	Borderlands	_45

Table 2 shows the arrangement of data and features in the dataset. There is a data section that is empty and looks incomplete in the Att4 feature. This condition needs to be corrected by using the replace missing value feature so that all empty data can be filled in.

2.3. Hyperparameter K-Nearest Neighbors (HKNN)

Hyperparameters are an important approach in machine learning because they can control algorithm behavior and train training and significantly affect machine learning models[25]. When using K-Nearest Neighbors (KNN), hyperparameters focus on parameters the user or researcher must define before the model training process begins. These hyperparameters are not learned or adjusted by the model itself during training, but they do affect the performance and characteristics of the KNN model. The following is the algorithm for the K-Nearest Neighbors Hyperparameter:

```
1.  Start
2.  Input a dataset (training data), a list of k values to evaluate (k_values), and a list of distance metrics
    to evaluate (distance_metrics).
3.  Initialize the best_accuracy variables with 0, best_k with 0, and best_distance_metric with an empty
    string.
4.  For each k value in k_values, do the following:
    a. For each distance metric in distance_metrics, do the following:
    b. Evaluate accuracy using cross-validation by calling the cross_validation function with dataset, k,
    and distance_metric arguments.
    c. If the resulting accuracy is higher than best_accuracy, update best_accuracy with that accuracy,
    best_k with k values, and best_distance_metric with distance metrics.
5.  Return best_k and best_distance_metric as the output of the algorithm.
6.  The cross_validation function accepts a dataset, k values, and a distance metric as arguments.
7.  Initialize a list of accuracy_scores.
    For each data_point in the dataset, perform the following steps:
    a. Separate data_point as test_point, and use another dataset as training_set.
    b. Use the KNN algorithm by calling the KNN function with the training_set, test_point, k, and
    distance_metric arguments.
    c. Compare the label predicted by KNN with the actual label from test_point.
    d. If the predicted labels are the same as the actual labels, add 1 to the accuracy_scores, otherwise
    add 0.
8.  Calculate accuracy by calculating the sum of 1's in accuracy_scores divided by the length of
    accuracy_scores.
9.  Return accuracy as the output of the cross_validation function.
10. The KNN function accepts training_set, test_point, k-value, and distance metric as arguments.
11. Initialize a list of distances.
12. For each training_point in the training_set, do the following:
    a. Calculate the distance between test_point and training_point using the given distance metric.
    b. Add the distance and label training_point to the distances list.
13. Sort the list of distances by distance ascending.
14. Get the nearest k labels from the distances list.
15. Use majority vote to choose the label that appears most often from k_nearest_labels.
16. Return the predicted label as the output of the KNN function.
17. Stop
```

2.4. Model Evaluation

Measurement of the performance of the resulting model is needed in machine learning. Model evaluation plays a special role in machine learning systems so that the user can evaluate the model's performance and determine how to improve it [26]. In this study, two ways were used to measure the performance of the resulting model, namely by using measurement techniques:

2.4.1 Accuracy

In data mining research, accuracy measures how well a data mining model or algorithm can correctly classify or predict results. Accuracy is often used as one of the most important evaluation metrics in data mining because the accuracy of the model or algorithm can affect the quality of the results. Accuracy can be formulated:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

2.4.2 Classification Error

Classification Error (CE) is used to evaluate the quality of the classification model. The smaller the classification error value, the better the classification model can classify data correctly. CE is usually calculated as the difference between the number of misclassified data and the total data in the dataset. Classification error is the inverse measure of accuracy in data mining. CE can be formulated as follows:

$$CE = (FP + FN) / (TP + TN + FP + FN) \quad (2)$$

Where:
TP is true positive, that is, the number of positive data correctly classified as positive.
TN is true negative, i.e., the amount of negative data correctly classified as negative.
FP is a false positive, i.e., the amount of negative data incorrectly classified as positive.
FN is a false negative, i.e., the number of positive data incorrectly classified as negative.

3. RESULTS AND DISCUSSION

Research on sentiment analysis on Twitter is often carried out using various approaches. This study used the KNN approach with hyperparameter tuning to carry out the classification to obtain high-accuracy results. Initially, this study used the KNN method by double preprocessing, namely by removing duplicates and replacing missing values. However, the accuracy results still did not achieve maximum results. However, after using KNN with the hyperparameter tuning approach, the level of accuracy increases. The research results before and after using the KNN method with hyperparameter tuning can be seen in Figure 2.

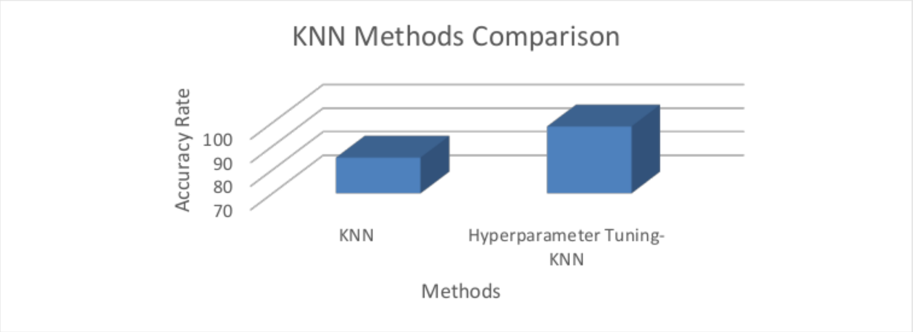


Figure 2. KNN research comparison

In Figure 2, it can be seen that the accuracy of the traditional KNN method reaches 85.15%. Otherwise, by using Hyperparameter Tuning-KNN, the accuracy rate increases to 98.37%. This research is also compared with several other studies conducted using the KNN method. The results of the research conducted, when compared with other studies, can be seen in Table 3.

Table 3. Comparison With Others KNN Research

Authors	Datasets	Accuracy (%)
Kaur [27]	sentiment analysis of Twitter data	86
Rezwanul [28]	Sentiment analysis on Twitter	84.32
Stephen [29]	amazon automotive product review dataset	85.53
Ours	**Twitter Sentiment Analysis**	**98.37**

Table 3 shows several comparisons of several studies conducted on sentiment analysis. Some researchers uses the KNN method to classify sentiment analysis datasets. Some researchers use the KNN method, and the results shows that the proposed method has the highest accuracy 98.37%. As for comparison with various other methods in Machine Learning and Deep Learning, it can be seen in Table 4.

Table 4. Comparison With Other Methods

Methods	Datasets	Accuracy (%)
Logistic Regression [30]	News Headlines Dataset for Sarcasm Detection	80
Random Forest [31]	tweets dataset with sarcastic or non-sarcastic labels	79.44
SVM[32]	Sarcastic tweets	74
Recurrent Neural Network [33]	Malayalam datasets	80
Proposed Method	**Twitter Sentiment Analysis**	**98.37**

The information in Table 4 shows the researcher also used various methods to conduct research in sentiment analysis. In table 4 shows the proposed method has an advantage in terms of high accuracy, namely at 98.37% and classification error 1.63% using KNN with hyperparameter tuning is better than other methods such as Linear Regression, Random Forest, SVM, and Recurrent Neural Network (RNN). The KNN with hyperparameter tuning method performs better than other methods. The performance results of the classification of the various methods can be seen in Figure 3.

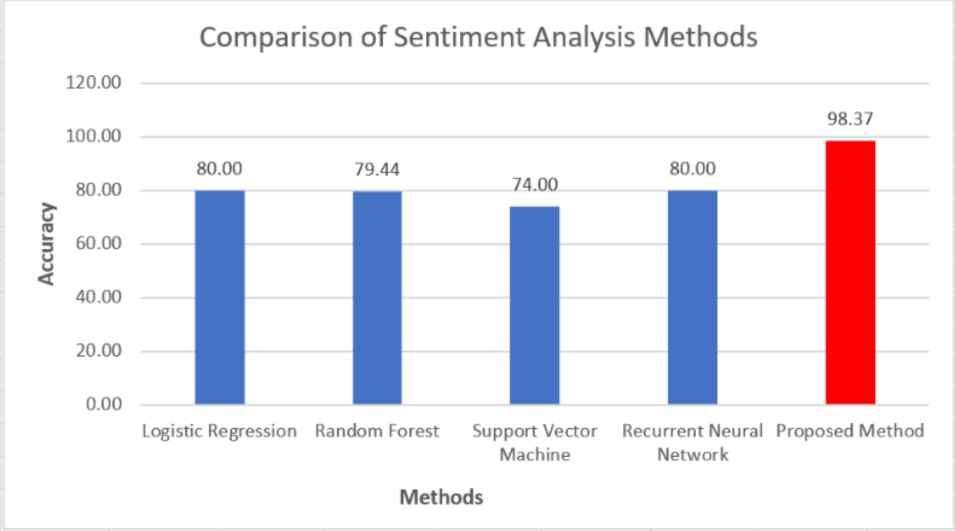


Figure 3. Sentiment Analysis Comparison Chart

KNN with hyperparameters is a robust approach. It can be shown in Figure 3 that KNN with the hyperparameter tuning method (the proposed method) has the best performance. The performance of KNN with hyperparameters as seen in the red bar graph, with an accuracy rate of 98.37%, is better than other methods.

4. CONCLUSION

Research in the field of data mining continues to develop on a variety of topics. One topic that is often researched is sentiment analysis. Various new methods have been created to improve the performance of the resulting model. The machine learning and deep learning approaches are popular methods that are often used in sentiment analysis. Still, no model produces performance with high accuracy even though it uses a deep learning approach. The KNN method, which is improved on the proposed hyperparameter tuning, produces a high accuracy of 98.37% for classifying sentiment analysis compared to other methods. In future research, Bidirectional Encoder Representations from Transformers (BERT) will be used in sentiment analysis research, so it is expected to produce a better model performance.

ACKNOWLEDGEMENTS

Authors express our gratitude to Universitas Stikubank for providing financing through incentives for scientific publications.

.

REFERENCES

[1] H. Raza, M. Faizan, A. Hamza, A. Mushtaq, and N. Akhtar, "Scientific text sentiment analysis using machine learning techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 12, pp. 157–165, 2019, doi: 10.14569/ijacsa.2019.0101222.

[2] K. K. Yusuf, E. Ogbuju, T. Abiodun, and F. Oladipo, "A Technical Review of the State-of-the-Art Methods in Aspect-Based Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 67–78, Feb. 2024, doi: 10.62411/jcta.9999.

[3] K. V. Ghag and K. Shah, "Conceptual Sentiment Analysis Model," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 4, p. 2358, Aug. 2018, doi: 10.11591/ijece.v8i4.pp2358-2366.

[4] A. Iorliam and J. A. Ingio, "A Comparative Analysis of Generative Artificial Intelligence Tools for Natural Language Processing," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 91–105, Feb. 2024, doi: 10.62411/jcta.9447.

[5] U. Singh, A. Saraswat, H. K. Azad, K. Abhishek, and S. Shitharth, "Towards improving e-commerce customer review analysis for sentiment detection," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-26432-3.

[6] N. K. Gondhi, Chaahat, E. Sharma, A. H. Alharbi, R. Verma, and M. A. Shah, "Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/3464524.

[7] T. Sinnasamy and N. N. A. Sjaif, "A Survey on Sentiment Analysis Approaches in e-Commerce," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 10, pp. 674–679, 2021, doi: 10.14569/IJACSA.2021.0121074.

[8] N. N. Wijaya, D. R. I. M. Setiadi, and A. R. Muslikh, "Music-Genre Classification using Bidirectional Long Short-Term Memory and Mel-Frequency Cepstral Coefficients," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 13–26, Jan. 2024, doi: 10.62411/jcta.9655.

[9] M. I. Akazue, I. A. Debekeme, A. E. Edje, C. Asuai, and U. J. Osame, "UNMASKING FRAUDSTERS: Ensemble Features Selection to Enhance Random Forest Fraud Detection," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 201–211, Dec. 2023, doi: 10.33633/jcta.v1i2.9462.

[10] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *J. Comput. Theor. Appl.*, vol. 2, no. 1, 2024, doi: 10.62411/jcta.10057.

[11] M. A. Fauzi, R. F. Nur Firmansyah, and T. Afirianto, "Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 16, no. 3, p. 1345, Jun. 2018, doi: 10.12928/telkomnika.v16i3.7751.

[12] P. Sasikala and L. Mary Immaculate Sheela, "Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00308-7.

[13] A. K. Bitto, M. H. I. Bijoy, M. S. Arman, I. Mahmud, A. Das, and J. Majumder, "Sentiment analysis from Bangladeshi food delivery startup based on user reviews using machine learning and deep learning," *Bull. Electr. Eng. Informatics*, vol. 12, no. 4, pp. 2282–2291, 2023, doi: 10.11591/eei.v12i4.4135.

[14] W. Huang, M. Lin, and Y. Wang, "Sentiment Analysis of Chinese E-Commerce Product Reviews Using ERNIE Word Embedding and Attention Mechanism," *Appl. Sci.*, vol. 12, no. 14, 2022, doi: 10.3390/app12147182.

[15] Y. Wen, Y. Liang, and X. Zhu, "Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China," *PLoS One*, vol. 18, no. 3 March, pp. 1–14, 2023, doi: 10.1371/journal.pone.0275382.

[16] M. Fu and L. Pan, "Sentiment Analysis of Tourist Scenic Spots Internet Comments Based on LSTM," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/5944954.

[17] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, p. 799, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14744.

[18] S. Sheikhi, M. T. Kheirabadi, and A. Bazzazi, "A Novel Scheme for Improving Accuracy of KNN Classification Algorithm Based on the New Weighting Technique and Stepwise Feature Selection," *J. Inf. Technol. Manag.*, vol. 12, no. 4, pp. 90–104, 2021, doi: 10.22059/jitm.2020.296305.2455.

[19] R. Kadry and O. Ismael, "A New Hybrid KNN Classification Approach based on Particle Swarm Optimization," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 291–296, 2020, doi: 10.14569/IJACSA.2020.0111137.

[20] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters," *J. Telecomunication, Electron. Comput. Eng.*, vol. 8, no. 12, pp. 59–65, 2016, Accessed: Oct. 17, 2019. [Online]. Available: http://journal.utem.edu.my/index.php/jtec/article/view/1436/947

[21] R. Wazirali, "An Improved Intrusion Detection System Based on KNN Hyperparameter Tuning and Cross-Validation," *Arab. J. Sci. Eng.*, vol. 45, no. 12, pp. 10859–10873, 2020, doi: 10.1007/s13369-020-04907-7.

[22] S. Ambesange, R. Nadagoudar, R. Uppin, V. Patil, S. Patil, and S. Patil, "Liver Diseases Prediction using KNN with Hyper Parameter Tuning Techniques," *Proc. B-HTC 2020 - 1st IEEE Bangalore Humanit. Technol. Conf.*, pp. 1–6, 2020, doi: 10.1109/B-

HTC50970.2020.9297949.

[23] H. Taherdoost, "Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects Hamed Taherdoost. Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Coll," *Int. J. Acad. Res. Manag.*, vol. 2021, no. 1, pp. 10–38, 2021.

[24] E. A. Felix and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," *IET Softw.*, vol. 13, no. 6, pp. 479–496, 2019, doi: 10.1049/iet-sen.2018.5193.

[25] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, 2019, doi: 10.11989/JEST.1674-862X.80904120.

[26] R. Fiebrink, P. R. Cook, and D. Trueman, "Human model evaluation in interactive supervised learning," *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 147–156, 2011, doi: 10.1145/1978942.1978965.

[27] S. Kaur, G. Sikka, and L. K. Awasthi, "Sentiment Analysis Approach Based on N-gram and KNN Classifier," *ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun.*, pp. 13–16, 2018, doi: 10.1109/ICSCCC.2018.8703350.

[28] M. Rezwanul, A. Ali, and A. Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 19–25, 2017, doi: 10.14569/ijacsa.2017.080603.

[29] K. V. K. Stephen, F. R. A. Al-Harthy, and M. T. Shaikh, "An Selection System For Automotive Sentiment Classification In Hadoop Using KNN Classifier," *ARPN J. Eng. Appl. Sci.*, vol. 15, no. 6, pp. 841–846, 2020.

[30] L. Novic, "A Machine Learning Approach to Text-Based Sarcasm Detection," pp. 1–24, 2022.

[31] A. G. Prasad, S. Sanjana, S. M. Bhat, and B. S. Harish, "Sentiment analysis for sarcasm detection on streaming short text data," *2017 2nd Int. Conf. Knowl. Eng. Appl. ICKEA 2017*, vol. 2017-Janua, no. 2009, pp. 1–5, 2017, doi: 10.1109/ICKEA.2017.8169892.

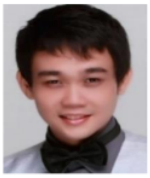
[32] N. Pawar and S. Bhingarkar, "Machine learning based sarcasm detection on twitter data," *Proc. 5th Int. Conf. Commun. Electron. Syst. ICCES 2020*, no. Icces, pp. 957–961, 2020, doi: 10.1109/ICCES48766.2020.09137924.

[33] M. Thomas and C. A. Latha, "Sentimental analysis using recurrent neural network," *Int. J. Eng. Technol.*, vol. 7, no. 2.27 Special Issue 27, pp. 88–92, 2018, doi: 10.14419/ijet.v7i2.27.12635.

BIOGRAPHIES OF AUTHORS

Kristiawan Nugroho works as a lecturer at Stikubank University's Faculty of Information Technology and Industry. In 2001, he graduated from Dian Nuswantoro University's Faculty of Computer Science with a bachelor's degree in information systems. Then, in 2007, he graduated from Dian Nuswantoro University with a master's degree in informatics engineering. In 2022, he also graduated from Dian Nuswantoro University Semarang with a doctorate in computer science with a focus on machine learning and artificial intelligence. He has conducted studies in sentiment analysis, speech recognition, and machine learning. His email address is kristiawan@edu.unisbank.ac.id.

Edy Winarno He is the rector of Stikubank University in addition to being a lecturer there. 2016, the Department of Computer Science awarded the recipient a Doctorate in Computer Science. At Stikubank University's Faculty of Information Technology and Industry, he presently holds the position of Associate Professor. Artificial intelligence, machine learning, computer vision, image processing, and security are among his areas of interest in research. His email address is edywin@edu.unisbank.ac.id.



De Rosal Ignatius Moses Setiadi Moses graduated with a bachelor's degree in 2010 from the Soegijaprana Catholic University in Semarang, Indonesia, and a master's degree from Dian Nuswantoro University in Semarang, Indonesia, in the same department. He works as a researcher and lecturer at the Dian Nuswantoro University's Faculty of Computer Science in Semarang, Indonesia. More than 138 peer-reviewed journal and conference papers that are indexed by Scopus have been written by him or with him. Artificial intelligence, watermarking, cryptography, and image steganography are among his areas of interest in research. His email address is moses@dsn.dinus.ac.id.

Prof Omar Farooq In 1992, Omar Farooq began his employment as a lecturer in the Department of Electronics Engineering at AMU Aligarh. Today, he holds the position of professor. In pursuit of his doctorate at Loughborough University in the UK, he was a Commonwealth Scholar from 1999 to 2002. In 2007–2008, he was granted a one-year UKIERI postdoctoral fellowship. He is interested in signal processing broadly, with a focus on speech recognition. He has guided nine researchers to their PhD graduation and wrote or co-authored more than 250 papers in conference proceedings and peer-reviewed academic journals. The Institute of Electrical and Electronics Engineers (IEEE, USA) has him as a Senior Member. Email correspondence with him can be sent to omar.farooq@amu.ac.in.

BEEI-03April2024-FinalPaper

ORIGINALITY REPORT

19%	15%	13%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ijai.iaescore.com Internet Source	4%
2	Submitted to Institute of Research & Postgraduate Studies, Universiti Kuala Lumpur Student Paper	2%
3	eprints.staffs.ac.uk Internet Source	1%
4	Ananthi, V.P., P. Balasubramaniam, and C.P. Lim. "Segmentation of gray scale image based on intuitionistic fuzzy sets constructed from several membership functions", Pattern Recognition, 2014. Publication	1%
5	eprints.uad.ac.id Internet Source	1%
6	Kristiawan Nugroho, Isworo Nugroho, De Rosal Igniatus Moses Setiadi, Omar Farooq. "Improving Indonesian multietnics speaker recognition using pitch shifting data	1%

	augmentation", IAES International Journal of Artificial Intelligence (IJ-AI), 2023 Publication	
7	0-www-mdpi-com.brum.beds.ac.uk Internet Source	<1 %
8	Hari Surrisyad, Wahyono -. "A Fast Military Object Recognition using Extreme Learning Approach on CNN", International Journal of Advanced Computer Science and Applications, 2020 Publication	<1 %
9	www.researchgate.net Internet Source	<1 %
10	publikasi.dinus.ac.id Internet Source	<1 %
11	www.irejournals.com Internet Source	<1 %
12	Submitted to Universidade do Porto Student Paper	<1 %
13	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2022 Publication	<1 %
14	Submitted to Daffodil International University Student Paper	<1 %
	garuda.kemdikbud.go.id	

15	Internet Source	<1 %
16	Achmad Nuruddin Safriandono, Muljono, Pujiono, Ruri Suko Safriandono. "Movie Sentiment Analysis Using Data Augmentation And LSTM-Recurrent Neural Network", 2023 International Seminar on Application for Technology of Information and Communication (iSemantic), 2023 Publication	<1 %
17	Submitted to Harrisburg University of Science and Technology Student Paper	<1 %
18	dokumen.pub Internet Source	<1 %
19	Submitted to SP Jain School of Global Management Student Paper	<1 %
20	Submitted to Universitas Dian Nuswantoro Student Paper	<1 %
21	doria.fi Internet Source	<1 %
22	beei.org Internet Source	<1 %
23	section.iaesonline.com Internet Source	<1 %

24	ujcontent.uj.ac.za Internet Source	<1 %
25	"The Proceedings of the 18th Annual Conference of China Electrotechnical Society", Springer Science and Business Media LLC, 2024 Publication	<1 %
26	Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko et al. "Review of Automatic Text Summarization Techniques & Methods", Journal of King Saud University - Computer and Information Sciences, 2020 Publication	<1 %
27	Andrew Giovanni Gozal, Hady Pranoto, Muhammad Fikri Hasani. "Sentiment analysis of the Indonesian community toward face-to-face learning during the Covid-19 pandemic", Procedia Computer Science, 2023 Publication	<1 %
28	ebin.pub Internet Source	<1 %
29	journals.ums.ac.id Internet Source	<1 %
30	www.adscientificindex.com Internet Source	<1 %

31	Gregor, D., S. Toral, T. Ariza, F. Barrero, R. Gregor, J. Rodas, and M. Arzamendia. "A methodology for structured ontology construction applied to intelligent transportation systems", Computer Standards & Interfaces, 2016. Publication	<1 %
32	Saeful Fahmi, Lia Purnamawati, Guruh Fajar Shidik, Muljono Muljono, Ahmad Zainul Fanani. "Sentiment Analysis of Student Review in Learning Management System Based on Sastrawi Stemmer and SVM-PSO", 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), 2020 Publication	<1 %
33	Shubham Shedekar, Sahil K. Shah, Vidya Kumbhar. "Enhancing E-Commerce Insights: Sentiment Analysis Using Machine Learning and Ensemble Techniques", 2023 International Conference on Integration of Computational Intelligent System (ICICIS), 2023 Publication	<1 %
34	export.arxiv.org Internet Source	<1 %
35	jurnal.iaii.or.id Internet Source	<1 %

36	ojs2.pnb.ac.id Internet Source	<1 %
37	telkomnika.uad.ac.id Internet Source	<1 %
38	ugspace.ug.edu.gh Internet Source	<1 %
39	www.mdpi.com Internet Source	<1 %
40	Haoyang Liu, Yaojin Lin, Chenxi Wang, Lei Guo, Jinkun Chen. "Semantic-gap-oriented feature selection in hierarchical classification learning", Information Sciences, 2023 Publication	<1 %
41	Mohamed Sharawy, Adel A. Shaltout, Omar El-Sayed Mohammed Youssef, Mahmoud A. Al-Ahmar, Naser Abdel-Rahim, Tole Sutikno. "Maximum allowable hp rating of 3-phase induction motor fed through a stand-alone constant V/f controlled DFIG via RSC", Bulletin of Electrical Engineering and Informatics, 2024 Publication	<1 %
42	"ACIT 2021 Conference Proceedings", 2021 22nd International Arab Conference on Information Technology (ACIT), 2021 Publication	<1 %

43	"Intelligent Computing", Springer Science and Business Media LLC, 2019 Publication	<1 %
44	"Recent Trends in Electronics and Communication", Springer Science and Business Media LLC, 2022 Publication	<1 %
45	archive.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		