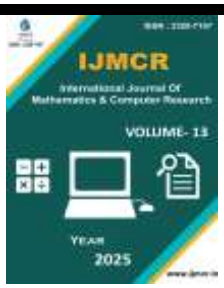
International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 13 Issue 07 July 2025, Page no. – 5359-5364

Index Copernicus ICV: 57.55, Impact Factor: 8.615

DOI: 10.47191/ijmcr/v13i7.01



A Comparative Study of Extreme Learning Machine Variants for Sentiment Analysis on Railway App Reviews

Nuki Pujiani Yosephine[1], Budi Warsito[2], Dinar Mutiara Kusumo Nugraheni[3]

[1]Master of Information Systems Program, Postgraduate School, Diponegoro University, Semarang, Indonesia

[2]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

[3]Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

ARTICLE INFO

Published Online:
01 July 2025

Corresponding Author:
Nuki Pujiani Yosephine

ABSTRACT

This study compares three Extreme Learning Machine (ELM) variants: ELM, Weighted-ELM (WELM), and Boosting Weighted-ELM (BWELM) for sentiment analysis of user reviews from Indonesian railway applications. Using TF-IDF for feature extraction and IndoBERT for labeling, the models were evaluated on an imbalanced dataset. ELM achieved 68.25% accuracy but struggled with minority classes. WELM improved performance to 72% by addressing class imbalance. BWELM, combining WELM and AdaBoost, achieved the best result with 78.25% accuracy, effectively handling imbalanced data. The findings highlight BWELM's potential for sentiment analysis in real-world, imbalanced datasets.

KEYWORDS: Extreme Learning Machine, Weighted-ELM, Boosting Weighted-ELM, Imbalanced Data, Sentiment Analysis

I. INTRODUCTION

The exponential growth of mobile applications has significantly increased the volume of user-generated content in the form of online reviews. These reviews often reflect users' satisfaction, expectations, and criticisms, making them a valuable source for sentiment analysis. Sentiment analysis, also known as opinion mining, is the computational study of people's opinions, sentiments, emotions, and attitudes expressed in textual data [1]. It plays a pivotal role in understanding public perception and improving service quality, especially in service-oriented industries such as transportation.

Railway services in Indonesia, particularly those managed through mobile platforms such as *Access by KAI* and *Tiket Kereta Api Online*, have witnessed widespread adoption by the public. Users frequently express their feedback and experiences on platforms like Google Play Store, providing an ideal corpus for analyzing customer sentiment [2]. Prior studies have demonstrated the importance of such sentiment analysis for enhancing service quality, as seen in analyses of transport and tourism apps [3].

Machine learning has become a cornerstone technique in sentiment analysis due to its ability to learn from data and generalize patterns. Among these, the Extreme Learning Machine (ELM) has gained attention for its fast learning speed and good generalization performance [4]. ELM is a single-hidden layer feedforward neural network that randomly assigns input weights and biases, and analytically determines output weights. While standard ELM offers efficient computation, its performance may degrade on imbalanced or complex datasets.

To address these limitations, variants such as Weighted-ELM (WELM) and Boosting Weighted-ELM (BWELM) have been proposed. WELM introduces cost-sensitive learning by assigning different weights to classes, thereby improving classification performance on imbalanced datasets [5]. BWELM further integrates boosting techniques like AdaBoost to iteratively enhance classification performance by focusing on difficult instances [6]. Such hybrid approaches have shown promising results in various domains including disease diagnosis and user behavior prediction [5].

This study presents a comparative evaluation of ELM, WELM, and BWELM using sentiment data collected from reviews of *Access by KAI* and *Tiket Kereta Api Online* apps. The reviews are preprocessed and merged into a single dataset, which is then labeled and analyzed. By applying and comparing these three models, this research aims to determine which approach yields the best accuracy in classifying user sentiments. The contributions of this study are as follows: (1) Construction of a unified sentiment dataset from two major

railway apps in Indonesia; (2) Implementation and evaluation of three ELM-based models: ELM, WELM, and BWELM; and (3) Comparative analysis of model performance in terms of accuracy and robustness on real-world review data.

II. METHODOLOGY

The methodology employed in this study to conduct sentiment analysis on user reviews of the railway applications *Access by KAI* and *Tiket Kereta Api Online*. The method in this research has a process flow as in Figure 1. The primary focus is on utilizing various Extreme Learning Machine (ELM) variants, including ELM, Weighted-ELM (WELM), and Boosting Weighted-ELM (BWELM) with AdaBoost, to analyze sentiment effectively.

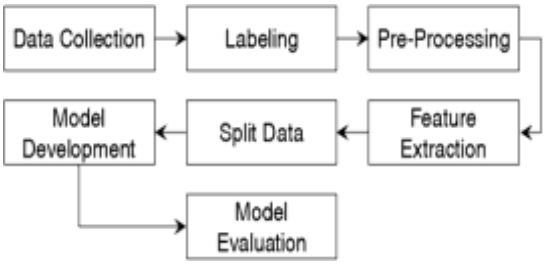


Figure 1. Research Process Flow

Data collection of user reviews for the *Access by KAI* and *Tiket Kereta Api Online* applications were collected using the Google Play Scraper library in Python [2], resulting in a dataset of 2000 samples for sentiment analysis. This use of user-generated content (UGC) offers valuable insights into customer experiences and preferences [3], [7].

Labeling assigns sentiment labels (positive, negative, or neutral) to each review, which is essential for model training. A pre-trained model, such as IndoBERT, was used to generate these labels based on the model's predictions [8].

Preprocessing involves transforming the raw text data into a format suitable for machine learning algorithms. The preprocessing steps include:

1) Cleansing involves removing irrelevant information from the text data, such as HTML tags, special characters, and punctuation. Using regular expressions to remove HTML tags and special characters, as well as punctuation from text data [9].
2) Normalization involves converting all text to lowercase to reduce dimensionality and improve model performance, by using the `lower()` function to convert all text to lowercase [10].
3) Tokenization involves splitting the text data into individual words or tokens, by using the `split()` function to split the text data into individual words [10].
4) Stopword removal involves eliminating common words that do not contribute to the sentiment, such as "and", "the", and "is". Using a list of stopwords to remove common words from the text data [9].
5) Stemming involves reducing words to their base or root form to ensure uniformity in analysis. Using a stemming algorithm, such as the Porter Stemmer, to reduce words to their base form [10].

Feature extraction transforms preprocessed text into a format suitable for machine learning. In this study, Term Frequency-Inverse Document Frequency (TF-IDF) was used to represent textual data, effectively capturing word importance across the dataset to improve sentiment classification [3]. The TF-IDF score for a term t in a document d is calculated as:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

where

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of tearms in document } d}$$

and

$$IDF(t) = log\left(\frac{N}{\text{Number of document containing term } t}\right)$$

when N as the total number of documents. The `TfidfVectorizer` from the `sklearn.feature_extraction.text` library was applied, limiting the features to the top 2000 terms to reduce dimensionality and enhance model performance.

The dataset was split into training and testing sets with an 80:20 ratio, and 10% of the training set was further reserved for validation to prevent overfitting and ensure generalization [8]. The `train_test_split` function from the `sklearn.model_selection` library was used for this process.

The model development in this study focuses on the construction and comparison of three classification models based on different ELM variants:

A. Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) is a machine learning algorithm designed for Single Layer Feedforward Neural Networks (SLFN). It significantly accelerates the training process by randomly initializing the weights and biases between the input and hidden layers, and then computing the output weights analytically using the Moore-Penrose pseudoinverse [4]. Given a training dataset $\{(x_i, t_i)\}_{i=1}^{N}$, the output of ELM is formulated as:

$$o_j = \sum_{i=1}^{L} \beta_i \cdot g\left(w_i \cdot x_j + b_i\right),\ j = 1, \dots, N \quad (2)$$

Where $g(\cdot)$ is the activation function, w_i and b_i are randomly initialized input weights and biases, and β_i are the output weights. The output weights are computed analytically as:

$$\beta = H^{\dagger}T \quad (3)$$

Where H is hidden layer output matrix, T is target output matrix, and $H^{\dagger}$ is the Moore-Penrose pseudoinverse of H. ELM offers high-speed training and good generalization performance; however, it performs sub-optimally when dealing with imbalanced datasets, which are frequently encountered in sentiment analysis tasks [11].

B. Weighted Extreme Learning Machine (WELM)

Weighted Extreme Learning Machine (WELM) is an enhancement of Extreme Learning Machine (ELM) specifically designed to handle class imbalance problems by incorporating a weighting mechanism that assigns greater importance to minority class samples during training [5]. The WELM formulation modifies the original ELM objective function as follows:

$$\min_{\beta} \| W^{1/2}(H\beta - T) \|^2 + \lambda \| \beta \|^2 \quad (4)$$

Where W is a diagonal matrix representing the weight of each training sample based on class distribution, λ is the regularization parameter, H is the hidden layer output matrix, and T is the target output matrix [5], [12]. This adjustment allows WELM to reduce bias toward majority classes and improve classification performance in real-world applications such as healthcare, social media sentiment analysis, and customer reviews [12].

C. Boosting Weighted-ELM (BWELM)

Boosting Weighted Extreme Learning Machine (BWELM) is an ensemble extension of Weighted Extreme Learning Machine (WELM) that integrates the AdaBoost algorithm to further improve classification performance, particularly for imbalanced datasets [6], [10], [13]. In BWELM, the AdaBoost algorithm operates by iteratively training multiple WELM classifiers on reweighted training data, where initial instance weights are set as $D_1(i) = \frac{1}{N}$ for all i. For each iteration $t = 1 \ldots T$, a WELM classifier $h_t(x)$ is trained using the current distribution D_1, and the error is calculated as:

$$\epsilon_t = \sum_i D_t(i) \cdot \mathrm{II}(h_t(x_i) \neq y_i) \quad (5)$$

The model weight is then computed by:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (6)$$

and the instance weights are updated as:

$$D_{t+1}(i) = \frac{D_t(i) \cdot exp(-\alpha_t y_i h_t(x_i))}{Z_t} \quad (7)$$

where Z_t is a normalization factor. The final ensemble classifier [6][10] is given by:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \quad (8)$$

The hyperparameters such as hidden layer size, regularization parameters, class weighting factors, and boosting iterations are tuned using grid search to optimize model performance [12] Throughout the tuning process, validation accuracy and loss are continuously monitored to identify the optimal configuration for each model, ensuring better generalization and minimizing overfitting risks [5].

After optimal hyperparameters are determined, the final configurations for ELM, WELM, and BWELM are fixed. Label encoders, feature extractors, and embedding layers are frozen to ensure consistent data representation, and the fully trained models are serialized for reproducible evaluation on unseen data [7].

The fixed models are evaluated on a separate test set, processed identically to the training data using text normalization, TF-IDF, and IndoBERT embeddings to ensure consistency. The ELM, WELM, and BWELM models generate predictions, which are assessed using accuracy, precision, recall, F1-score, and confusion matrix to comprehensively evaluate performance across sentiment classes [1][3].

Overall the model design is constructed to address the cpmplexity of sentiment classification where non-linearity [1], class imbalance, and noisy data are prevalent [14], particularly in real-world datasets from user-generated content [7]. The evaluation compares ELM, WELM, and BWELM based on their ability to handle class imbalance [5], the effects of weighting and boosting, and generalization on unseen data [13], providing insights into each model's strengths and limitations in sentiment analysis [15].

III. RESULT AND DISCUSSION

A. Model Training and Validation

During training and validation, each of the models: ELM, WELM, and BWELM was trained using 80% of the total dataset, with 10% reserved for validation. The ELM model, using randomly initialized input weights and Moore-Penrose pseudoinverse for output weights, achieved 70.42% training and 72% validation accuracy. However, it struggled with neutral and positive classes, showing respectively weighted average precision 0.61, recall 0.70, and F1-score 0.62.WELM, incorporating class-weighted loss functions to address imbalance, improved performance with 74.03% training and 75% validation accuracy, achieving better balance in neutral (recall 0.32) and positive (recall 0.74) classes, and weighted averages of 0.74 for precision, recall, and f1-score. BWELM, combinig WELM with AdaBoost, delivered the best result, achieving 81.25% training and 80% validation accuracy, with recalss of 0.94 (negative), 0.48 (neutral), and 0.79 (positive), and weighted average precision, recall, and F1-score of 0.81, demonstrating the effectiveness of boosting in handling minority class misclassification. The performance variations of each model during training and validation are presented in Table 1.

Tabel 1. Comparison on Training and Validation Dataset

Model	*Training Accuracy*	*Validation Accuracy*
ELM	69.37%	68.13%
WELM	85.14%	64.38%
BWELM	97.29%	80.00%

B. Model Testing

The final evaluation was conducted on a completely unseen testing dataset consisting of 400 samples. Each model was assessed based on its confusion matrix, precision, recall, F1-score, and overall accuracy.

1. ELM

The ELM model achieved an overall testing accuracy of 68.25%. The confusion matrix on Fig.2, revealed that ELM performed well in classifying negative sentiment (precision 0.71, recall 0.94), but failed to correctly classify the neutral class (recall 0.00) and struggled on positive class samples (recall 0.15, precision 0.36). This imbalance demonstrates that while ELM is efficient for majority class prediction, its random weight initialization and lack of compensation for imbalanced data result in poor generalization for minority classes.

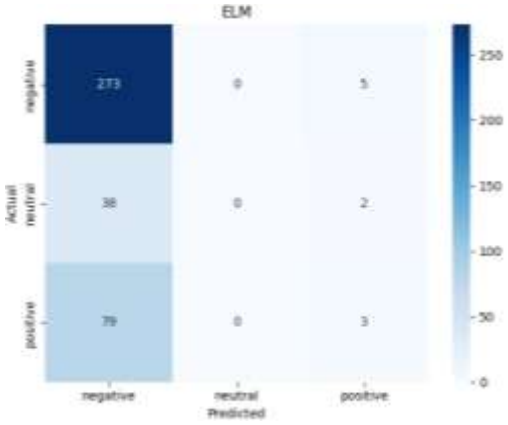


Figure 2. ELM Confusion Matrix

2. *WELM*

The WELM model improved performance across all classess, yielding an overall accuracy of 72.00%. Its neutral class recall increased to 0.50, while the positive class recall rose to 0.80, with a corresponding precision of 0.61. The class-weighted loss function enabled WELM to better capture the minority classes, particularly neutral and positive sentiments, compared to standar ELM.

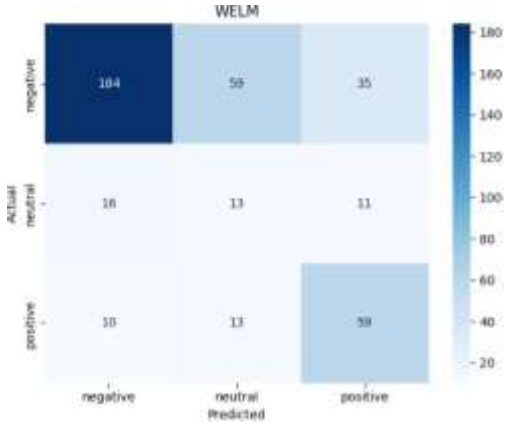


Figure 3. WELM Confusion Matrix

3. *BWELM*

BWELM achieved the highest overall accuracy of 78.25%, showing considerable improvement over both ELM and WELM. The confusion matrix on Fig.4, reflected balanced precision and recall across all classes. For the negative class, recall reached 0.93; for the neutral class, recall was 0.50; and for the positive class, recall reached 0.68 with higher precision compared to both ELM and WELM. The AdaBoost-based boosting mechanism allowed the BWELM model to progressively focus on difficult-to-classify instances, resulting in superior generalization across imbalanced classes.

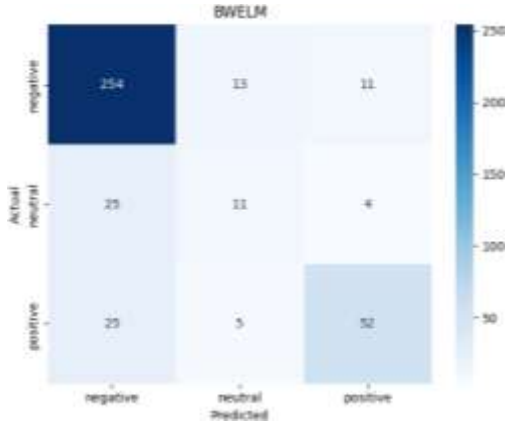


Figure 4. BWLEM Confusion Matrix

C. *Model Comparison*

A comparative analysis of the three models: ELM, WELM, and BWELM is presented in Table 2.

Tabel 2. Model Accuracy Comparison on Testing Dataset

Model	*Testing Accuracy*
ELM	68.25%
WELM	72.00%
BWELM	78.25%

Based on the comparison in Table 1, we can see the process of each model in revealing different strengths and weaknesses for each approach in handling sentiment classification on an imbalanced data set. The visualization of the comparison of the three models is based on Fig.5 and Fig.6.

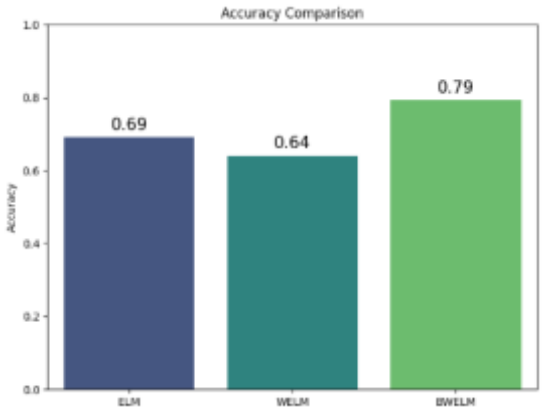


Figure 5. Model Accuracy Comparison

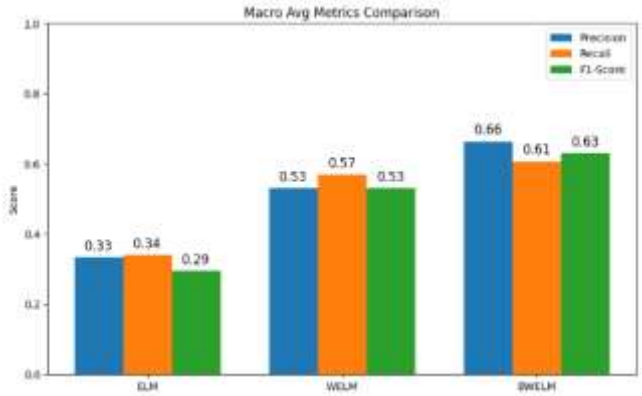


Figure 6. Comparison of Macro Average Model Metrics

The ELM model demonstrated its primary strength in terms of simplicity and computational efficiency. Since ELM does not require iterative training and utilizes random weight initialization with analytical computation of output weights, it is capable of extremely fast training even on large datasets. This makes ELM highly attractive for situations where computational resources are limited or rapid deployment is needed. However, ELM exhibited significant weaknesses when applied to imbalanced sentiment data, as it failed to adequately classify minority classes. Its performance was heavily skewed towards the majority (negative) class, with poor sensitivity to both neutral and positive classes. This limitation stems from ELM's inability to incorporate any mechanism for imbalance handling or error focusing during training.

The WELM model introduced improvements by incorporating class weighting into the training process. This allowed the model to better account for class distribution differences, resulting in improved recall and F1-scores for the minority classes, particularly for neutral and positive sentiments. WELM's strength lies in its capability to adjust the influence of each class on the learning process, thus reducing the dominance of the majority class during model training. Nevertheless, WELM still showed certain weaknesses, as its performance gains were limited in the absence of more dynamic learning adjustments. The model continued to struggle with instances located near class boundaries or noisy data points, as it lacked adaptive mechanisms to further refine learning from previously misclassified samples.

The BWELM model demonstrated superior performance by combining the class-weighted learning of WELM with the adaptive reweighting mechanism of AdaBoost. Its strength lies in its ability to progressively focus on difficult-to-classify instances by increasing the weight of misclassified samples in subsequent learning iterations. This adaptive boosting mechanism significantly improved BWELM's sensitivity across all sentiment classes, resulting in a more balanced and robust classification performance. As a consequence, BWELM achieved the highest overall accuracy among the three models and demonstrated the most consistent precision and recall across positive, neutral, and negative sentiments. However, this improved performance comes at the cost of increased model complexity and computational time due to the ensemble structure and iterative boosting process, which may require more processing resources compared to ELM and WELM.

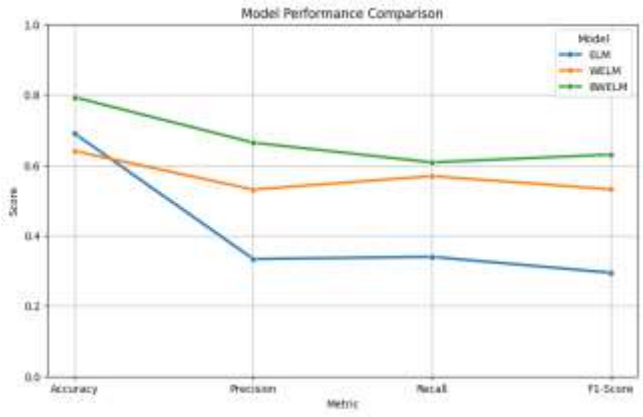


Figure 7. Model Performance Comparison

In summary, while ELM offers simplicity, and WELM effectively addresses imbalance through weighting, BWELM emerges as the most effective solution by combining both weighting and boosting to achieve comprehensive performance improvements in sentiment classification.

C. Discussion

The results clearly demonstrate that model enhancements addressing both class imbalance and hard-to-classify instances significantly improve sentiment classification performance. ELM's simplicity provides efficient training but is insufficient for imbalanced sentiment data commonly found in user reviews. The addition of class weighting in WELM shows meaningful improvement by allowing the model to pay more attention to minority classes. However, BWELM's integration of AdaBoost further improves the model's capability by iteratively adjusting instance weights based on previous misclassifications, resulting in the most balanced and robust performance across all sentiment categories. These findings are consistent with previous studies where hybrid boosting and weighting strategies outperformed traditional single classifiers in sentiment analysis tasks [5][13][15].

IV. CONCLUSION

A comparative analysis was conducted to evaluate the performance of three Extreme Learning Machine (ELM) based models: ELM, Weighted-ELM (WELM), and Boosting Weighted-ELM (BWELM), for sentiment analysis on railway application reviews. The experimental results demonstrated that while ELM provides a simple and computationally efficient solution, it struggles to accurately classify minority sentiment classes due to its inability to address class imbalance. Incorporating class weights in WELM improved performance by assigning higher penalties to misclassified minority class samples, leading to better balance across sentiment categories, particularly for the neutral and positive classes.

However, BWELM consistently outperformed both ELM and WELM by integrating class weighting with adaptive boosting through the AdaBoost algorithm. The boosting mechanism allowed BWELM to focus learning on misclassified instances

across multiple iterations, resulting in superior generalization and balanced performance across all sentiment classes. The BWELM model achieved the highest overall testing accuracy and exhibited the most stable precision, recall, and F1-scores for all sentiment categories, making it the most effective approach for sentiment analysis on the given imbalanced dataset.

In conclusion, the integration of both class weighting and boosting provides a significant advantage in handling real-world sentiment analysis tasks where class imbalance is often inevitable. The BWELM model, therefore, can be recommended as a highly effective solution for sentiment classification problems, particularly in analyzing user-generated reviews from railway applications or other service-based industries with similar data characteristics. Future research may focus on further improving model robustness by exploring advanced ensemble techniques or incorporating deep learning-based feature extraction for enhanced sentiment representation.

REFERENCES

1. M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif Intell Rev, vol. 55, no. 7, pp. 5731–5780,Oct.2022,doi:10.1007/s10462-022-10144-1.
2. P. D. Atika, H. a, and F. N. Khasanah, "SENTIMENT ANALYSIS OF KAI ACCESS APPLICATION USING THE DEEP NEURAL NETWORK METHOD," IJARCCE, vol. 10, no. 12, Dec. 2021, doi: 10.17148/ijarcce.2021.101205.
3. K. Puh and M. B. Babac, "Predicting sentiment and rating of tourist reviews using machine learning," Journal of Hospitality and Tourism Insights, vol. 6, no. 3, pp. 1188–1204, Jun. 2023, doi: 10.1108/JHTI-02-2022-0078.
4. J. Wang, S. Lu, S. H. Wang, and Y. D. Zhang, "A review on extreme learning machine," Multimed Tools Appl, vol. 81, no. 29, pp. 41611–41660, Dec. 2022, doi: 10.1007/s11042-021-11007-7.
5. S. Priya and Dr. R. Manavalan, "A hybrid Classifier using Weighted ELM with Biogeography Based Optimization for Hepatitis Diagnosis," in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), IEEE, 2019, pp. 1463–1467.
6. X. Feng, "Research of sentiment analysis based on adaboost algorithm," in Proceedings - 2019 International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2019, Institute of Electrical and Electronics Engineers Inc., Nov. 2019, pp. 279–282. doi: 10.1109/MLBDBI48998.2019.00062.
7. W. Chen, Z. Xu, X. Zheng, Q. Yu, and Y. Luo, "Research on Sentiment Classification of Online Travel Review Text," Applied Sciences (Switzerland), vol. 10, no. 15, pp. 1–21, Aug. 2020, doi: 10.3390/APP10155275.
8. J. Hidayat, S. M. Honova, V. Pangesa, C. A. Setiawan, I. H. Parmonangan, and Diana, "Sentiment Analysis of Skincare Product Reviews in Indonesian Language using IndoBERT and LSTM," in Proceeding - IEEE 9th Information Technology International Seminar, ITIS 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ITIS59651.2023.10420222.
9. R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naive Bayes," in Proceedings of 2019 International Conference on Information & Communication Technology and Systems (ICTS), IEEE, 2019, pp. 49–54.
10. M. Rathi, A. Malik, D. Varghney, R. Dharma, and S. Mediratta, "Sentiment Analysis of Tweets using Machine Learning Approach," in 2018 Eleventh International Conference on Contemporary Computing (IC3), IEEE, 2018.
11. M. Kaur, D. Das, and S. P. Mishra, "Survey and Evaluation of Extreme Learning Machine on TF-IDF Feature for Sentiment Analysis," in Proceedings - 2022 International Conference on Machine Learning, Computer Systems and Security, MLCSS 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 247–252. doi: 10.1109/MLCSS57186.2022.00053.
12. O. K. Utomo, N. Surantha, S. M. Isa, and B. Soewito, "Automatic sleep stage classification using weighted ELM and PSO on imbalanced data from single lead ECG," Procedia Comput Sci, vol. 157, pp. 321–328, 2019, doi: 10.1016/j.procs.2019.08.173.
13. B. S. Raghuwanshi and S. Shukla, "Class-specific cost-sensitive boosting weighted ELM for class imbalance learning," Memet Comput, vol. 11, no. 3, pp. 263–283, Sep. 2019, doi: 10.1007/s12293-018-0267-4.
14. A. Jazuli, Widowati, and R. Kusumaningrum, "Aspect-based sentiment analysis on student reviews using the Indo-Bert base model," in E3S Web of Conferences, EDP Sciences, Nov. 2023. doi: 10.1051/e3sconf/202344802004.
15. F. Daneshfar and S. J. Kabudian, "Speech Emotion Recognition Using Multi-Layer Sparse Auto-Encoder Extreme Learning Machine and Spectral/Spectro-Temporal Features with New Weighting Method for Data Imbalance," in ICCKE 2021 - 11th International Conference on Computer Engineering and Knowledge, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 419–423. doi: 10.1109/ICCKE54056.2021.9721524.