Improving sentiment classification accuracy through information gain feature selection and hyperparameter tuning in random forest model

Doni Abdul Fatah[1*], *Muhammad* Ali Syakur[1], *Yudha* Dwi Putra Negara[1], *Nasrulloh* Nasrulloh[2]

[1]Department of Informatics Engineering, Faculty of Engineering, Universitas Trunodjoyo Madura, Bangkalan, Indonesia
[2]Department of Sharia Economics, Faculty of Islamic Studies, Universitas Trunodjoyo Madura, Bangkalan, Indonesia

Abstract. This study aims to enhance the accuracy of sentiment classification for Madura coastal tourism reviews using Random Forests optimised via Information Gain (IG) feature selection and hyperparameter tuning. The research contribution is an efficient hybrid framework that improves classification performance by integrating data-driven feature-relevance analysis and parameter optimisation techniques. The methodology involves several stages, including data scraping from online travel platforms, text preprocessing (cleaning, tokenisation, and stemming), TF-IDF feature extraction, feature selection using IG thresholds, and model optimisation using GridSearchCV. Experimental results show that the combination of Information Gain and hyperparameter tuning significantly increases the Random Forest model's accuracy. At an IG threshold of 0.0002, the model achieved the highest accuracy of 89.69%, while the 0.0004 threshold provided nearly similar accuracy (89.18%) with fewer features, making it more efficient. These findings indicate that careful feature selection and parameter tuning can improve model generalisation and efficiency. In conclusion, this study contributes to advancing sentiment classification in tourism data analytics through an optimised Random Forest model and provides insights for future research in data-driven modelling for decision support in the tourism sector.

1 Introduction

Tourism is a strategic sector that plays an important role in driving regional and national economic growth through travel, recreation, and the exploration of cultural and natural potential [1]. In Indonesia, including the Madura region, tourism has become one of the leading sectors contributing to local economic development, job creation, cultural preservation, and supporting infrastructure development [2].

Various strategies have been implemented to enhance the competitiveness of tourist destinations, including improving facilities, optimising digital promotion, and strengthening tourism-based creative economies to improve community welfare [3][4]. One of the main contributors to Regional Original Revenue (PAD) is coastal tourism, which serves as a flagship destination that continues to grow through promotional activities and infrastructure development [5][6].

In the digital era, social media and travel platforms play a significant role in tourism promotion. Through reviews, photos, and videos shared online, tourists can influence public perceptions of destinations both positively and negatively [7][8][9]. These reviews have become a vital source of data for analysing tourist satisfaction and perceptions of destinations [10]. However, challenges arise when review texts contain spelling or typographical errors, which can reduce the accuracy of text data analysis. Such errors hinder Natural Language Processing (NLP) and affect sentiment analysis results that serve as a basis for decision-making [11][12][13].

In the context of Madura's coastal tourism, sentiment analysis of tourist reviews plays an essential role in assisting the government and destination managers to improve service quality and formulate effective marketing strategies [14][15]. Therefore, an analytical approach that can enhance the accuracy of sentiment classification models is required.

In recent years, sentiment analysis has been increasingly recognised as part of the broader field of computational science and data-driven modelling, which is closely related to physics and applied mathematics through statistical learning and optimisation methods. In this scientific context, algorithms such as Random Forest and Information Gain implement probabilistic models and entropy-based feature evaluation concepts that are also fundamental in physical information theory.

Text Mining is the process of extracting information from unstructured text data to enable large-scale data analysis. Natural Language Processing (NLP) is used to identify and extract opinions from text, transforming textual content into quantitative data for decision-making [16]. Sentiment analysis is one of the main applications of text mining, aiming to classify opinions in text into positive, negative, or neutral categories [17]. The initial stage of sentiment analysis is Natural

* Corresponding author: doni.fatah@trunojoyo.ac.id

Language Preprocessing, which includes a series of techniques to prepare text data for analysis [18].

One of the most widely used classification methods in sentiment analysis is the Random Forest Classifier (RF) [19][20]. This algorithm delivers strong performance compared to other classification methods due to its ensemble learning, which combines multiple decision trees. However, RF also has weaknesses in terms of accuracy and data stability, especially when too many or irrelevant features are included [21][22]. Therefore, feature extraction and selection techniques are necessary to improve the model's accuracy and stability [23][24][25].

One of the approaches to enhance accuracy is by applying the Information Gain (IG) method for feature selection and Hyperparameter Tuning on the Random Forest (RF) model [26][27]. Information Gain, based on the principle of entropy reduction, helps identify the most relevant features for sentiment categories, reducing noise and improving model efficiency [28][29]. Meanwhile, hyperparameter tuning optimises key parameters of the RF algorithm—such as the number of trees, maximum depth, and number of selected features—to achieve optimal classification performance [30].

Accurate feature correction and selection are crucial stages in text mining and NLP because both play a central role in extracting meaningful information from unstructured text [31]. In sentiment analysis, NLP is used to identify opinions and represent them in quantitative data that supports decision-making [32]. The Random Forest algorithm is known for its strong performance and generalisation ability compared to other classification methods; however, it still faces challenges related to model stability and sensitivity to imbalanced data [33]. Therefore, applying Information Gain for feature selection and Hyperparameter Tuning is essential to enhance accuracy, robustness, and consistency. Additionally, implementing N-Gram and TF-IDF methods during feature extraction helps capture word patterns and relationships, ensuring that only the most influential terms for sentiment classification are retained[34].

In this study, the improvement of sentiment classification accuracy is achieved through the combination of Information Gain feature selection and hyperparameter tuning on the Random Forest model. The tuning process aims to optimise parameters such as the number of trees, maximum depth, and number of features selected for each decision tree, leading to optimal classification performance. The combination of these two approaches is expected to improve the performance of the RF model in analysing tourist reviews and produce more accurate data to support decision-making for the development of the tourism sector in Madura.

2 Methods

This study applies a quantitative experimental approach comprising several main stages: data collection, preprocessing, feature extraction, feature selection, modelling, and evaluation. The overall research workflow is illustrated in Fig. 1.

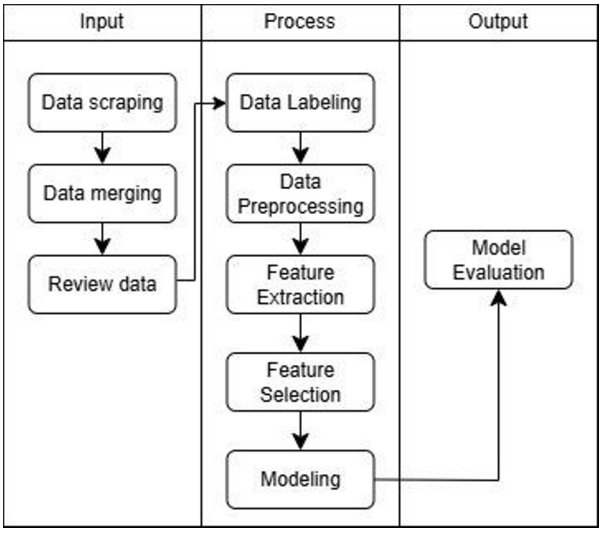

Fig. 1. Research flowchart showing the input–process–output stages.

The Input stage includes acquiring tourism review data; the Process stage involves text preprocessing, feature extraction, feature selection, and model development; and the Output stage produces the final evaluation results of the sentiment classification model.

2.1 Data collection

The initial stage of this study involved collecting data in the form of tourist reviews and comments on coastal tourism in Madura. The data were obtained from various social media and travel platforms using a data scraping technique [35].

The scraping process used Python libraries such as BeautifulSoup and Selenium to extract textual content from posts and comment sections related to Madura's coastal tourism. After merging data from each platform, a total of 1,634 review entries were collected for sentiment analysis.

2.2 Data labelling

The scraped data were then manually labelled to categorise opinions into two sentiment classes: positive and negative. Three different annotators conducted the labelling process to maintain consistency and data reliability.

If differences of opinion occurred during labelling, a majority voting method was used, with the most frequently chosen label as the final result. A linguistics expert subsequently validated the labelling outcomes to ensure contextual and semantic accuracy. As a result, 966 reviews were labelled as positive and 668 as negative.

2.3 Pre-processing

Pre-processing was conducted to structure the collected reviews and facilitate subsequent data analysis [36]. This stage included several steps as follows:

1. Case Folding – converting all text to lowercase for uniformity.
2. Cleaning – removing non-alphabetic characters, punctuation marks, numbers, URLs, emojis, and special symbols.
3. Tokenising – splitting text into smaller units (tokens).
4. Normalisation – converting informal or non-standard words into their standard forms.
5. Filtering (Stopword Removal) – removing common words that do not contribute significant meaning.
6. Stemming – reducing words to their root forms using an Indonesian stemming algorithm.

This preprocessing step produces a clean text corpus that is ready for feature extraction [37].

2.4 Feature extraction

Feature extraction aims to transform text into numerical form using the Term Frequency–Inverse Document Frequency (TF-IDF) technique [38]. TF-IDF calculates the weight of each term based on its frequency in a document relative to its frequency across the entire corpus.

The processed dataset is transformed into a numerical vector matrix by computing term frequency (TF) using Equation (2), followed by inverse document frequency (IDF) using Equation (3), and then calculating the TF-IDF value with Equation (4). Finally, v-norm weighting is applied using Equation (5) [38].

2.5 Feature selection

After feature extraction, feature selection was performed using the Information Gain (IG) method to identify the most relevant features associated with sentiment categories. Information Gain measures the reduction in entropy, allowing the selection of features that contribute the most to classification accuracy [39][40].

Different IG threshold values (0.0002, 0.0004, 0.0006, 0.0008) were tested to evaluate their effect on the Random Forest model's performance. Features with IG values above the threshold were retained, while those below were removed to reduce dimensionality and noise.

2.6 Modelling and hyperparameter tuning

Before model training, the dataset was split into training (80%) and test (20%) sets. The Random Forest algorithm was used as the primary classifier due to its robustness and ensemble-based performance [41]. To optimize model accuracy, GridSearchCV was applied for hyperparameter tuning, exploring the best combinations of parameters such as *n_estimators*, *max_depth*, *min_samples_split*, and *min_samples_leaf* [42].

The optimal parameters obtained from the training phase were then used to construct the final sentiment classification model.

2.7 Model evaluation

Model evaluation aimed to measure the classification performance of the trained Random Forest model. Evaluation metrics included accuracy, precision, recall, and F1-score, all derived from the confusion matrix [43][44][45].

This evaluation process ensures the model's reliability by assessing both correctness and the balance between positive and negative sentiment predictions.

3 Result and discussion

3.1 Feature extraction

The feature extraction stage employed the Term Frequency–Inverse Document Frequency (TF-IDF) technique, which converts the stemmed text into numerical vectors representing term importance within the corpus [38]. The implementation using the Scikit-learn library produced a sparse matrix with weighted term frequencies, as illustrated in Table 1.

Tabel 1. TF-IDF results.

D	**father**	**event**	**…**	**tourism**	**form**
1	0.384898	0	…	0	0
2	0	0	…	0	0
3	0	0	…	0	0
…	…	…	…	…	…
1633	0		…	0	0
1634	0	0	…	0	0

The TF-IDF weighting enabled the model to distinguish important words for sentiment classification by assigning higher weights to terms that appear frequently in a document but less frequently in other documents. This process ensured that only the most discriminative words were retained for feature selection using Information Gain.

3.2 Feature selection

Feature selection was conducted using the Information Gain (IG) method to determine which features contributed most to classification accuracy. Features with IG values above the threshold were selected, while the rest were removed to reduce noise and dimensionality.

Four threshold values (0.0002, 0.0004, 0.0006, and 0.0008) were tested to evaluate model sensitivity to feature relevance. As shown in Table 2, a lower threshold retained more features but did not always yield better accuracy.

Table 2. Information gain results and selected features.

Feature No	Feature	*Information Gain*
1	father	0.002033
2	event	0.002325
…	…	…
2030	tourism	0.004660
2031	form	0.002107

This indicates that selecting a moderate threshold value (0.0004) provides an efficient trade-off between the number of features and classification accuracy, reducing computational cost without significant loss of performance.

3.3 Model optimisation and evaluation

Before model evaluation, the dataset was split into 80% for training and 20% for testing. The Random Forest model was trained with optimal hyperparameters obtained using GridSearchCV: n_estimators = 145, max_depth = None, min_samples_split = 5, and min_samples_leaf = 1.

The model achieved its highest accuracy of 89.69% at an IG threshold of 0.0002, while the 0.0004 threshold produced nearly identical accuracy (89.18%) with fewer features, as presented in Table 3. The confusion matrix in Fig. 2 demonstrates that the model achieved balanced classification between positive and negative classes.

Table 3. Evaluation matrix results of the testing scenario.

Threshold (Accuracy)	Total Features with IG	Class	Precision	*Recall*	*F1-Score*
0,0008 (88,14%)	578	Positive	90%	85%	87%
		Negative	87%	91%	89%
0,0006 (88,14%)	1277	Positive	87%	89%	88%
		Negative	90%	87%	88%
0,0004 (89,18%)	1845	Positive	92%	85%	88%
		Negative	87%	93%	90%
0,0002 (89,69%)	1965	Positive	90%	88%	89%
		Negative	89%	91%	90%

These results confirm that feature selection and hyperparameter tuning effectively improve classification accuracy and model stability.

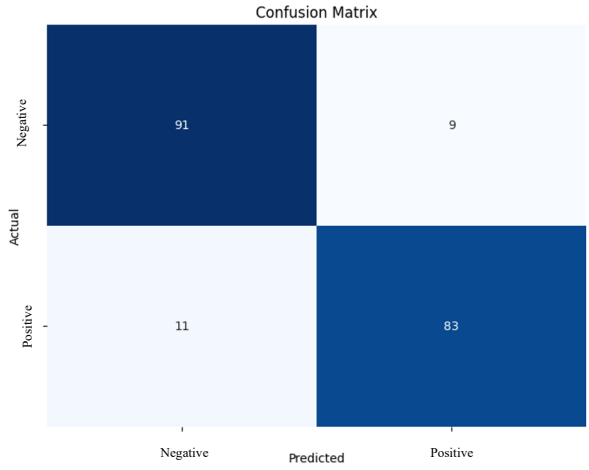

Fig. 2. Confusion matrix of the best trial scenario.

3.4 Comparison with previous studies

Compared to prior studies using traditional models such as Naïve Bayes and Support Vector Machine (SVM), the proposed Information Gain–Random Forest (IG-RF) approach shows better performance in sentiment classification. For instance, the Naïve Bayes models used by Jonnala et al. [18] and Alsemaree et al. [43] achieved accuracies of 82%-85%, whereas our optimised IG-RF model reached 89.69%.

Similarly, the work of Kanwar et al. [27] demonstrated the effectiveness of Random Forest combined with grid search for optimisation in environmental modelling, supporting our finding that hyperparameter tuning contributes significantly to improved model generalisation. Moreover, the integration of Information Gain for feature selection, as applied in studies by Zhou and Xue [28] And Dhanka et al. [29], also proved effective in reducing dimensionality while maintaining model performance. These comparisons confirm that combining entropy-based feature selection and parameter tuning provides substantial benefits for text-based sentiment analysis.

3.5 Implications and explanation of findings

The main finding of this study is that the Information Gain threshold directly influences model efficiency and interpretability. A smaller threshold retains more features, which may slightly improve accuracy but increases computational load, while a moderate threshold (0.0004) maintains nearly equivalent accuracy with fewer features. The integration of GridSearchCV tuning enhances the model's adaptability to diverse data distributions, thereby improving generalisation capability.

3.6 Strengths and limitations

The strength of this study lies in combining Information Gain and hyperparameter tuning to systematically optimise the Random Forest model, which effectively balances accuracy and efficiency. Furthermore, the inclusion of a real-world dataset (tourist reviews)

demonstrates the practical applicability of data-driven sentiment analysis in social sciences.

However, the study has certain limitations. The dataset is limited to binary sentiment categories (positive/negative) and does not include neutral expressions. Moreover, spelling correction and semantic-level preprocessing were not implemented, which could further enhance accuracy in future research. Future studies may integrate deep learning architectures such as BERT or LSTM to capture contextual information and improve model performance.

4 Conclusion

This study successfully improved sentiment classification accuracy by integrating Information Gain (IG) feature selection and hyperparameter tuning into the Random Forest (RF) model. The experimental results demonstrated that applying IG thresholds and tuning parameters such as n_estimators, max_depth, min_samples_split, and min_samples_leaf effectively optimised the model's accuracy and stability. The highest accuracy of 89.69% was achieved at an IG threshold of 0.0002, while a slightly lower threshold of 0.0004 yielded comparable accuracy (89.18%) with fewer features, indicating that feature efficiency can be achieved without sacrificing model performance.

The theoretical contribution of this study lies in demonstrating the effectiveness of entropy-based feature selection combined with parameter optimisation as a robust framework for text-based sentiment analysis. This finding supports the principle that reducing entropy through Information Gain aligns with improved classification consistency and generalisation in ensemble models. From an applied perspective, the proposed model provides a reliable analytical framework for understanding public opinion—particularly in tourism-related datasets—where accuracy and computational efficiency are essential.

However, this research still has limitations. The dataset used was restricted to binary sentiment categories (positive and negative), excluding neutral or mixed opinions that may provide deeper insights. In addition, semantic-level preprocessing and spelling correction were not implemented, which may affect the model's ability to handle noisy text inputs.

For future work, it is recommended to incorporate semantic-based preprocessing methods such as word embeddings (Word2Vec, FastText) or contextual deep learning architectures (LSTM, BERT) to capture meaning and contextual relationships more effectively. Further studies could also expand the sentiment classes and explore cross-domain datasets to enhance model generalisation across different application areas.

Sincere thanks to the Directorate General of Higher Education, Science, and Technology, Ministry of Education, Science, and Technology of the Republic of Indonesia, and to Universitas Trunodjoyo Madura, supported by the Institute for Research and Community Service (LPPM UTM), for the funding support provided for this research.

References

1. E. J. Ijatuyi, K. Yessoufou, H. O. Patrick, Sustainable tourism and green space: exploring how green spaces and natural attractions contribute to local tourism economies and revenue generation in Gauteng province. Discov. Sustain. **6**, 1, 236 (2025). https://doi.org/10.1007/s43621-025-00958-8
2. N. Bakalo, V. Makhovka, I. Krekoten, A. Glebova, S. Kulakova, Local tourism as financial and economic development driver of the community: Management aspect. World Dev. Perspect. **38**, 100693 (2025). https://doi.org/10.1016/j.wdp.2025.100693
3. L. Judijanto, Inclusive and Sustainable Growth Through Culture-Based and Creative Economy Tourism: A Review. Multitech J. Sci. Technol. **2**, 9, 667–688 (2025). https://doi.org/10.59890/mjst.v2i9.71
4. I. Nurhayati et al., Community-Based Halal Tourism and Information Digitalization: Sustainable Tourism Analysis. Tour. Hosp. **6**, 3 (2025). https://doi.org/10.3390/tourhosp6030148
5. J. Pan, H. Yang, Z. Wang, B. Peng, S. Li, Optimizing Sustainable Tourism: A Multi-Objective Framework for Juneau and Beyond. Sustainability **17**, 16 (2025). https://doi.org/10.3390/su17167344
6. I. F. Pasaribu, T. D. Hapsari, G. Y. Prasetya, M. Sciences, Strategy For Developing Blue Economic Potential Through Sustainable Tourism In The Coastal Area Of Semarang City. J. Teknol. Perikan. dan Kelaut. **16**, 2, 173–188 (2025). https://doi.org/10.24319/jtpk.16.173-188
7. S. Priatmoko, A. F. Rahmat, Digging up Rural Community-based tourism (CBT) in developing country, Indonesia's framework finding. Geoj. Tour. Geosites **61**, 3, 1420–1429 (2025). https://doi.org/10.30892/gtg.61302-1512
8. S. Shin, T. Kim, S. Hlee, C. Koo, Destination Advertising on YouTube: Effects of Native Advertising and Comment Management on Tourist Perception. J. Hosp. & Tour. Res. **49**, 4, 798–815 (2025). https://doi.org/10.1177/10963480231194689
9. L. Fang, Y. Liu, G. Li, From Residents to Tourists: Influence of Online Review Negativity Dominance on Tourism Public Opinion. In HCI in Business, Government and Organizations. (2025), 3–17
10. X. Huang, S. Chelliah, Attributes Influencing Tourist Satisfaction: Sentiment Analysis and Topic Modeling of Online Reviews. J. China Tour. Res. **21**, 3, 819–838 (2025). https://doi.org/10.1080/19388160.2024.2440323
11. N. A. Sharma, A. B. M. S. Ali, M. A. Kabir, A review of sentiment analysis: tasks, applications, and deep learning techniques. Int. J. Data Sci. Anal. **19**, 3, 351–388 (2025). https://doi.org/10.1007/s41060-024-00594-x

12. E. W. T. Ngai, A. K. H. Lui, B. C. W. Kei, Natural language processing in government applications: a literature review and a case analysis. Ind. Manag. Data Syst. **125**, 6, 2067–2104 (2025). https://doi.org/10.1108/IMDS-07-2024-0711
13. S. Lu, News editorial decision optimization and intelligent assistance based on social media and natural language processing. Serv. Oriented Comput. Appl. (2025). https://doi.org/10.1007/s11761-025-00457-8
14. S. Arifin, M. Azinuddin, A. P. Mat Som, A. Ibrahim, M. H. Hanafiah, Collaborative communication for sustainable tourism in Asia: a case study from Madura Island. Worldw. Hosp. Tour. Themes **17**, 3, 413–421 (2025). https://doi.org/10.1108/WHATT-01-2025-0042
15. S. Tunca, Exploring visitor sentiment trends at Alanya Cleopatra Beach using natural language processing techniques: insights from online reviews. Tour. Recreat. Res. **0**, 0, 1–20 (2025). https://doi.org/10.1080/02508281.2025.2503992
16. J. Köckritz, B. İlgen, C. Cohrdes, G. Hattab, Current applications and future directions in natural language processing for news media and mental health. Sci. Rep. **15**, 1, 32532 (2025). https://doi.org/10.1038/s41598-025-18413-z
17. P. López-Úbeda, T. Martín-Noguerol, A. Luna, Natural language processing and LLMs in liver imaging: a practical review of clinical applications. Abdom. Radiol. (2025). https://doi.org/10.1007/s00261-025-05127-z
18. N. S. Jonnala et al., Leveraging hybrid model for accurate sentiment analysis of Twitter data. Sci. Rep. **15**, 1, 24438 (2025). https://doi.org/10.1038/s41598-025-09794-2
19. R. Alhejaili, Machine Learning Approaches for Sentiment Analysis on Social Media. In AI-Driven: Social Media Analytics and Cybersecurity (2025), 21–43. https://doi.org/10.1007/978-3-031-80334-5_2
20. M. A. Bouke, O. I. Alramli, A. Abdullah, XAIRF-WFP: a novel XAI-based random forest classifier for advanced email spam detection. Int. J. Inf. Secur. **24**, 1, 5 (2024). https://doi.org/10.1007/s10207-024-00920-1
21. N. S. K. M. K. Tirumanadham et al., Optimizing Lung Cancer Prediction Models: A Hybrid Methodology Using GWO and Random Forest. In Enabling Person-Centric Healthcare (2025), 59–77. https://doi.org/10.1007/978-3-031-82516-3_3
22. E. Faiella et al., Promising Results About the Possibility to Identify Prostate Cancer Patients Employing a Random Forest Classifier. Diagnostics **15**, 4 (2025). https://doi.org/10.3390/diagnostics15040421
23. A. Yaqoob et al., SGA-Driven feature selection and random forest classification for enhanced breast cancer diagnosis. Sci. Rep. **15**, 1, 10944 (2025). https://doi.org/10.1038/s41598-025-95786-1
24. S. Wali, Y. A. Farrukh, I. Khan, Explainable AI and Random Forest based reliable intrusion detection system. Comput. Secur. **157**, 104542 (2025). https://doi.org/10.1016/j.cose.2025.104542
25. E. Caputi et al., Comparison of Tree Typologies Mapping Using Random Forest. Remote Sens. **17**, 3 (2025). https://doi.org/10.3390/rs17030356
26. W. Hu et al., Intelligent irrigation strategy model for farmland using dung beetle optimization-random forest algorithms. Agric. Water Manag. **317**, 109653 (2025). https://doi.org/10.1016/j.agwat.2025.109653
27. M. Kanwar, B. Pokharel, S. Lim, A new random forest method for landslide susceptibility mapping. Int. J. Environ. Sci. Technol. **22**, 11, 10635–10650 (2025). https://doi.org/10.1007/s13762-024-06310-3
28. Q. Zhou, F. Xue, Automatic information gain-guided convergence for refining building design parameters. Build. Environ. **275**, 112788 (2025). https://doi.org/10.1016/j.buildenv.2025.112788
29. S. Dhanka et al., Advancements in Hybrid Machine Learning Models for Biomedical Disease Classification. Arch. Comput. Methods Eng. (2025). https://doi.org/10.1007/s11831-025-10309-5
30. K. Saneep et al., State of charge estimation of lithium-ion batteries using PSO optimized random forest algorithm. J. Energy Storage **114**, 115879 (2025). https://doi.org/10.1016/j.est.2025.115879
31. O. M. Alyasiri et al., Text classification based on optimization feature selection methods: a review. Multimed. Tools Appl. **84**, 15, 14187–14233 (2025). https://doi.org/10.1007/s11042-024-19769-6
32. R. Nasfi, G. de Tré, A. Bronselaer, Improving Data Cleaning by Learning From Unstructured Textual Data. IEEE Access **13**, 36470–36491 (2025). https://doi.org/10.1109/ACCESS.2025.3543953
33. B. Priya Kamath et al., Comprehensive Analysis of Word Embedding Models and Feature Vector Design. IEEE Access **13**, 25239–25255 (2025). https://doi.org/10.1109/ACCESS.2025.3536631
34. E. J. Hadi, M. F. Ibrahim, A. I. Mohammed, Towards News Classification: A Machine Learning Approach. J. Inf. & Knowl. Manag. **24**, 05, 2550045 (2025). https://doi.org/10.1142/S0219649225500455
35. M. Basirati, A. Laachach, Clustering sustainable tourism destinations through Instagram photo analysis. Tour. Rev. (2025). https://doi.org/10.1108/TR-11-2024-0987
36. S. Khalighi et al., Evaluating the impact of data pre-processing methods on classification of ATR-FTIR spectra. Fuel **376**, 132701 (2024). https://doi.org/10.1016/j.fuel.2024.132701
37. B. Dayananda et al., Pre-processing Applied to Instrumental Data in Analytical Chemistry. Crit.

Rev. Anal. Chem. **54**, 8, 2745–2753 (2024). https://doi.org/10.1080/10408347.2023.2199864

38. Z. Labd et al., Text classification supervised algorithms with TF-IDF and GloVe: a comparative study. Int. J. Electr. Comput. Eng. **14**, 1, 589–599 (2024). https://doi.org/10.11591/ijece.v14i1.pp589-599
39. I. Emmanuel, Y. Sun, Z. Wang, A machine learning-based credit risk prediction engine system. J. Big Data **11**, 1, 23 (2024). https://doi.org/10.1186/s40537-024-00882-0
40. G. Obaido et al., Detecting Thyroid Disease Using Filter-Based Selection and Stacking Ensemble. IEEE Access **12**, 89098–89112 (2024). https://doi.org/10.1109/ACCESS.2024.3418974
41. M. Burukanli, N. Yumuşak, StackGridCov: stacking ensemble model with GridSearchCV for mutation prediction. Neural Comput. Appl. **36**, 35, 22379–22401 (2024). https://doi.org/10.1007/s00521-024-10428-3
42. A. Ahmed Bilal et al., Children's Sentiment Analysis From Texts Using Random Forest. IEEE Access **12**, 70089–70104 (2024). https://doi.org/10.1109/ACCESS.2024.3400992
43. O. Alsemaree et al., Sentiment analysis of Arabic social media texts. Heliyon **10**, 9 (2024). https://doi.org/10.1016/j.heliyon.2024.e27863
44. Y. Rimal et al., Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest. Sci. Rep. **15**, 1, 13444 (2025). https://doi.org/10.1038/s41598-025-93675-1
45. M. M. Mijwil, M. M. Mijwil, Comparative Analysis of Machine Learning Algorithms for Diabetes Classification. Baghdad Sci. J. **21**, 5, 24 (2024). https://doi.org/10.21123/bsj.2023.9010