

Comparative Analysis of Machine Learning Algorithms for Sentiment Classification on Restaurant Reviews Using TF-IDF Feature Extraction

Ronald Manela Putra Jaya

Department of Information Systems, Faculty of Computer Science, Universitas Katolik Soegijapranata,
Semarang, Central Java, 50234, Indonesia
e-mail: 23.n1.0041@student.unika.ac.id

Diterima
04-08-2026

Direvisi
11-08-2026

Disetujui
12-08-2026

Abstract: Online platforms for restaurant reviews are producing more user-written feedback than ever, and buried in that feedback is a lot of useful signal about food quality, service, and the overall dining experience. Making sense of it by hand isn't realistic anymore, which is part of why automated sentiment classification keeps getting attention. This paper compares four supervised classifiers for that task Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine (SVM) applied to multiclass sentiment classification of restaurant reviews. Out of 10,000 customer reviews collected, 9,948 were usable after cleaning. Each review was assigned to one of three sentiment classes (negative, neutral, positive) based on its numerical rating. Before modeling, the text went through case folding, removal of punctuation and numbers, stopword removal, and stemming, then TF-IDF vectorization followed by variance-based feature selection. An 80:20 split was used for training and testing, and the models were scored on accuracy, precision, recall, and macro-averaged F1. Random Forest came out ahead on accuracy, but SVM had the better macro F1 a gap worth noting, since it shows accuracy by itself can give a misleading picture once the dataset is imbalanced. The takeaway is that both the evaluation metric and the algorithm need to be chosen carefully when sentiment classes aren't evenly distributed, as is often the case with restaurant reviews.

Keywords: Sentiment Analysis, TF-IDF, Machine Learning, Class Imbalance, Restaurant Reviews

Abstrak: Pertumbuhan pesat platform ulasan restoran daring telah menghasilkan volume umpan balik tekstual yang besar dari pengguna, yang merangkum beragam opini pelanggan mengenai kualitas makanan, pelayanan, dan pengalaman bersantap secara keseluruhan. Penelitian ini melakukan analisis komparatif terhadap empat algoritme klasifikasi machine learning terawasi *Logistic Regression*, *Naïve Bayes*, *Random Forest*, dan *Support Vector Machine* (SVM) untuk klasifikasi sentimen multikelas pada ulasan restoran. Dataset yang terdiri dari 10.000 ulasan pelanggan dikumpulkan, dan setelah pembersihan data, 9.948 ulasan dipertahankan serta dikategorikan ke dalam tiga kelas sentimen negatif, netral, dan positif berdasarkan rating numeriknya. Data tekstual melalui *pipeline* praproses sistematis yang terdiri dari *case folding*, penghapusan tanda baca dan angka, penghapusan stopwords, dan stemming, dilanjutkan dengan pembobotan *Term Frequency Inverse Document Frequency* (TF-IDF) dan seleksi fitur berbasis varians. Dataset dibagi dengan rasio 80:20, dan kinerja model dievaluasi menggunakan *Accuracy*, *Precision*, *Recall*, dan *F1-score* rata-rata makro. Hasil penelitian menunjukkan bahwa Random Forest memperoleh akurasi keseluruhan tertinggi, sedangkan SVM memperoleh *F1-score* makro tertinggi, yang menunjukkan bahwa akurasi saja dapat menyesatkan pada dataset yang tidak seimbang. Temuan ini menegaskan pentingnya pemilihan metrik evaluasi dan algoritma yang tepat dalam menangani tugas klasifikasi sentimen yang tidak seimbang pada domain ulasan restoran.

Kata kunci: Analisis Sentimen, TF-IDF, Machine Learning, Ketidakseimbangan Kelas, Ulasan Restoran

I. INTRODUCTION

Digital platforms have become central to how the restaurant industry operates, since customers now routinely share their dining experiences in public through written reviews as well as star ratings. Buried in that text is a lot of detail: how the food tasted, whether service was quick or slow, what the ambience felt like, how satisfied the customer was overall. For restaurant owners trying to improve and for diners deciding where to eat, that detail is genuinely useful (Mehedi et al., 2025). The problem is volume once reviews start piling up in the thousands, reading through

them one by one stops being realistic. Sentiment analysis (sometimes called opinion mining) exists largely to solve that problem, and it's become one of the go-to techniques in Natural Language Processing (NLP) for pulling subjective opinions out of raw text automatically. At its core, the task is just figuring out a text's polarity positive, neutral, or negative(Mao et al., 2024). Applied to restaurants, this means a business can track how customers actually feel at scale, instead of relying on a handful of reviews someone happened to read, and use that to inform decisions. Sentiment classification used to learn heavily on lexicon-based approaches, like the unsupervised semantic-orientation method(Wankhade et al., 2022). These methods are easy to interpret, but they run into trouble when it comes to scalability and adapting to context. As labelled datasets became more available, supervised machine learning approaches started taking over the field. were among the first to apply Naïve Bayes, Maximum Entropy, and Support Vector Machines to review classification, and saw considerable gains over the lexicon-based baselines. Later work extended these techniques to product and restaurant reviews, and found that the choice of feature representation whether Bag-of-Words, n-grams, or TF-IDF has a real effect on model performance(Kamath et al., 2025).

Before any classifier can do its job, the text has to be turned into numbers that's what feature extraction handles in this pipeline. TF-IDF (Term Frequency Inverse Document Frequency) is still a common choice for this, largely because it's cheap to compute and does a decent job of highlighting words that actually carry meaning while down weighting the common but empty ones(Haiman, 1998). It also holds up empirically: recent work has found TF-IDF outperforming plain Bag-of-Words counts across several classical classifiers(Barron, 2025) on restaurant-review data specifically. A number of supervised learning algorithms have been extensively applied in sentiment classification research. Logistic Regression is recognized for its simplicity and reliable, interpretable performance on linear classification problems(Zahra et al., 2026).

Naïve Bayes is popular for its computational efficiency and probabilistic framework, which suits high-dimensional textual data well, even though its conditional independence assumption rarely holds up in practice(Pajila et al., 2023). Random Forest has repeatedly shown strong classification performance thanks to its ensemble of trees mechanism. Naïve Bayes is widely adopted owing to its computational efficiency and probabilistic framework, making it suitable for high-dimensional textual data, although its conditional-independence assumption rarely holds in practice(Kohli, 2025). Random Forest has consistently demonstrated strong classification performance through its ensemble of trees mechanism. Meanwhile, Support Vector Machine (SVM) has repeatedly shown competitive performance in text classification tasks by constructing optimal separating hyperplanes in high dimensional, sparse feature spaces such as those produced by TF-IDF(Yarsasi & Iskoko, 2025). More recently, transformer-based architectures such as BERT(Gillioz et al., 2020) and multilingual, aspect focused models(Rahman et al., 2025) have been explored for sentiment tasks, but classical models remain attractive for small to medium, domain-specific datasets because of their lower computational cost and greater interpretability(Zahra et al., 2026) Fuzzy-logic enhanced approaches have also been proposed to better capture sentiment intensity in restaurant-review contexts(Rokhva et al., 2025). Beyond classification, the information obtained from data analysis can also serve as supporting information in decision-making processes, particularly when decisions involve multiple alternatives and evaluation criteria. In such contexts, Decision Support Systems (DSS) and Multi-Criteria Decision-Making (MCDM) methods can be employed to systematically evaluate alternatives and produce objective rankings. One of the MCDM approaches, Multi-Objective Optimization on the basis of Ratio Analysis (MOOSRA), provides a structured mechanism for evaluating alternatives based on multiple criteria and their relative importance. This approach is therefore relevant to selection problems in which candidates must be evaluated comprehensively based on criteria such as leadership ability, communication skills, dedication, and organizational experience(Pradana et al., 2025).

Model evaluation in sentiment classification is commonly conducted using Accuracy, Precision, Recall, and F1-score derived from the confusion matrix. However, when the class distribution is imbalanced, accuracy alone may present a misleadingly favorable picture of model

performance, since a classifier may achieve high accuracy simply by favoring the majority class. Macro-averaged F1-score, which computes the F1-score for each class independently and averages them with equal weight, is therefore considered more appropriate for imbalanced multiclass classification tasks (Farhadpour et al., 2024).

Even though sentiment analysis of online reviews has been widely studied, comparing multiple classical machine learning algorithms under the same experimental conditions specifically on restaurant review data is still worth investigating, especially when the dataset has class imbalance across sentiment categories (Al-Qudah et al., 2025). With that in mind, this paper carries out a systematic comparison of four supervised machine learning algorithms Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine for multiclass sentiment classification of restaurant reviews. The collected review texts go through a structured preprocessing pipeline, TF-IDF feature extraction, variance-based feature selection, and are evaluated using Accuracy, Precision, Recall, and F1-score.

The main contributions of this paper are as follows: (1) it puts forward an end-to-end sentiment analysis pipeline built for restaurant review data, including standardized text preprocessing and TF-IDF-based feature extraction; (2) it gives an empirical comparison of four widely used supervised machine learning algorithms under the same experimental conditions; (3) it identifies the most effective classification algorithm for sentiment analysis of restaurant reviews, judged across multiple evaluation metrics, including one suited to imbalanced class distributions; and (4) it offers practical insights for researchers and restaurant industry practitioners on choosing suitable machine learning models for customer opinion mining.

II. RESEARCH METHOD

1. Research Framework

The approach here is quantitative and experimental: four supervised algorithms are run on the same multiclass sentiment classification task and compared directly. The full workflow moves through dataset collection, data cleaning, sentiment labelling, text preprocessing, TF-IDF feature extraction, feature selection, dataset partitioning, model training, and finally performance evaluation.

2. Dataset

The dataset has 10,000 restaurant reviews to start, each one carrying a reviewer name, restaurant name, the review text itself, and a rating on a 1-5 scale. Cleaning was needed because some ratings weren't numeric or were simply missing, and a handful of reviews had no text at all those got dropped. What was left after cleaning was 9,948 usable reviews. To make multiclass sentiment classification possible, the numerical ratings were converted into three sentiment categories: reviews rated below 3 were labelled Negative, those rated exactly 3 were labelled Neutral, and those rated above 3 were labelled Positive.

3. Text Processing

A series of preprocessing steps were applied to clean up the textual data before feature extraction, run in sequence: case folding, URL removal, punctuation removal, numerical character removal, stopword removal, and stemming.

4. Feature Extraction

After preprocessing, all review documents were converted into numerical representations using TF-IDF vectorization, with the *Quanteda* package in R used to build the document-feature matrix (DFM) and compute TF-IDF weights for each term.

5. Feature Selection

Since TF-IDF vectorization produces a high-dimensional, sparse feature space, variance-based feature selection was applied to cut dimensionality while keeping the most informative features. The variance of each TF-IDF feature was computed, and only the features with the

highest variance values were kept for model training.

6. Dataset Partitioning

The processed dataset was split into training and testing subsets using an 80:20 hold-out strategy, with a fixed random seed of 123 to ensure the results could be reproduced.

7. Classification Models

Four supervised machine learning algorithms were implemented under the same experimental conditions: (1) Logistic Regression, run as a multinomial classifier to handle the three sentiment categories; (2) Naïve Bayes, run as a probabilistic classifier that estimates posterior class probabilities; (3) Random Forest, run as an ensemble of decision trees aggregated through majority voting; and (4) Support Vector Machine, run using a linear kernel function.

8. Performance Evaluation

Performance for each model was read off a confusion matrix built from its predictions on the held out test set, and four metrics were pulled from that: Accuracy, Precision, Recall, and macro-averaged F1-score. Because the sentiment classes in this dataset aren't evenly distributed, macro F1-score was given priority over the others as the main indicator of which classifier actually held up best(Farhadpour et al., 2024).

III. RESULTS

1. Dataset Overview

9,948 valid reviews remained once cleaning was done. Table 1 breaks down how those reviews landed across the three sentiment classes after rating-based labelling.

Table 1. Distribution of Sentiment Labels

Sentiment	Number of Reviews	Percentage
Negative	2,447	24.60%
Neutral	1,191	11.97%
Positive	6,310	63.43%
Total	9,948	100%

Figure 1 lays out that distribution visually positive reviews dominate the dataset by a wide margin, and the neutral class is barely represented in comparison.

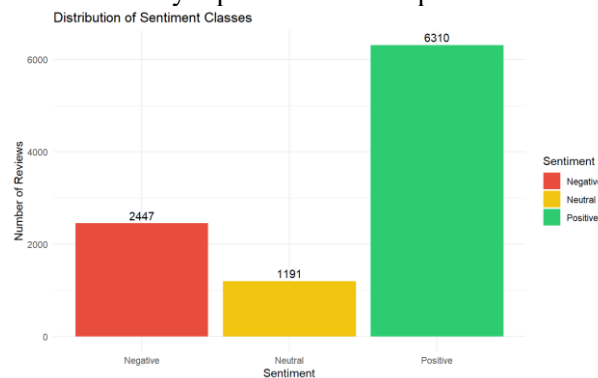


Figure 1. Distribution of Sentiment Classes

Figure 2 goes back a step and shows the raw rating distribution before any labelling happened, and it tells basically the same story: most customers rated their experience quite highly.

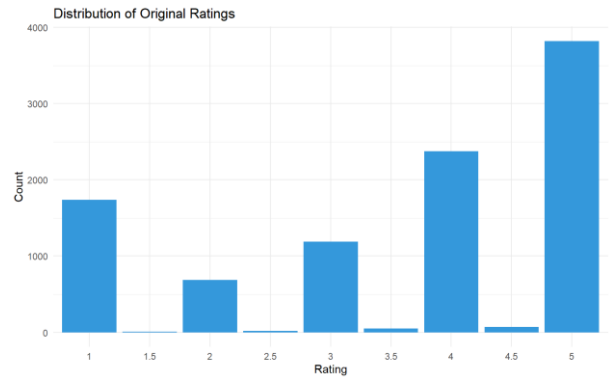


Figure 2. Distribution of Original Ratings

Review length, shown in Figure 3, skews short overall the median comes out to just 34 words.

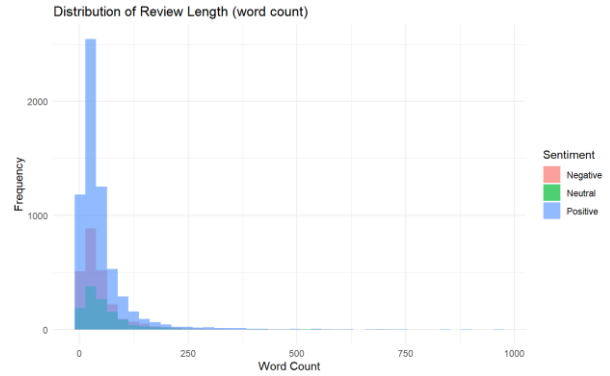


Figure 3. Distribution of Review Length (Word Count)

2. Text Preprocessing Result

The preprocessing steps were applied one after another: case folding, then URL and punctuation removal, stopwords elimination, and stemming last. Table 2 shows a few examples of what raw review text looked like before and after this process.

Table 2. Example of Text Preprocessing

Before	After
“The ambience was good, food was quite good...”	ambiencc good food quit good saturday lunch...
“Ambience is too good for a pleasant evening...”	ambiencc good pleasant even servic prompt food good...
“A must try.. great food great ambience...”	must tri great food great ambiencc thnx servic...

The preprocessing stage cut down textual noise while keeping the meaningful information intact. Stemming in particular pulled morphological word variants together into their root forms, resulting in a more compact and consistent vocabulary for feature extraction.

3. TF-IDF Feature Extarction and Feature Selection

Starting out, the document-feature matrix built from the preprocessed reviews had 12,834 unique terms a lot to work with. Trimming out terms that showed up fewer than five times across the whole corpus cut that down to 3,377. From there, variance-based feature selection kept only the 1,000 highest-variance features for training. Table 3 walks through each stage of that reduction.

Table 3 Feature Reduction Process

Stage	Number of Features
Before Trimming	12,834
After Trimming	3,377

After Feature Selection 1,000

4. Experimental Setup

An 80:20 hold-out split gave 7,959 reviews for training and 1,989 for testing. To keep the comparison fair, all four classifiers were trained on that exact same TF-IDF feature set nothing different between them except the algorithm it self.

5. Logistic Regression Results

The multinomial Logistic Regression classifier reached an overall accuracy of 78.13%, with macro-averaged Precision, Recall, and F1-score of 65.07%, 64.52%, and 64.75%.

Table 4 Performance of Logistic Regression

Metric	Value
Accuracy	78.13%
Precision	65.07%
Recall	64.52%
F1-score	64.75%

6. Naïve Bayes Results

The Naïve Bayes classifier turned in the lowest performance among the four algorithms tested, with accuracy of just 38.91%, despite a relatively higher macro Precision of 47.16%.

Table 5 Performance of Naïve Bayes

Metric	Value
Accuracy	38.91%
Precision	47.16%
Recall	47.89%
F1-score	35.57%

7. Random Forest Results

Random Forest reached the highest overall accuracy among the classifiers tested, at 81.45%. The confusion matrix (Figure 4) shows the model had a markedly low recall of just 3.78% for the Neutral class.

Table 6 Performance of Random Forest

Metric	Value
Accuracy	81.45%
Precision	73.07%
Recall	59.80%
F1-score	58.56%

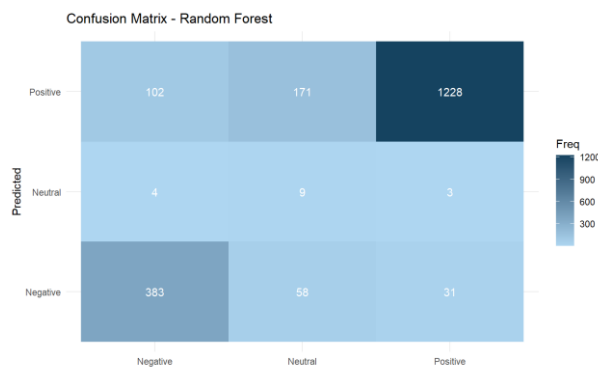


Figure 4 Confusion Matrix Heatmap Random Forest

8. Support Vector Machine Results

The linear-kernel SVM classifier reached an accuracy of 78.53%, along with the highest

macro-averaged F1-score among all models tested, at 65.03%.

Table 7. Performance of Support Vector Machine

Metric	Value
Accuracy	78.53%
Precision	66.03%
Recall	64.23%
F1-score	65.03%

9. Comparative Summary

Table 8 sums up the performance of all four classifiers across the four evaluation metrics, and Figure 5 gives a visual comparison of Accuracy, Precision, Recall, and F1-score across the models.

Table 8 Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	78.13%	65.07%	64.52%	64.75%
Naïve Bayes	38.91%	47.16%	47.89%	35.57%
Random Forest	81.45%	73.07%	59.80%	58.56%
Support Vector Machine	78.53%	66.03%	64.23%	65.03%

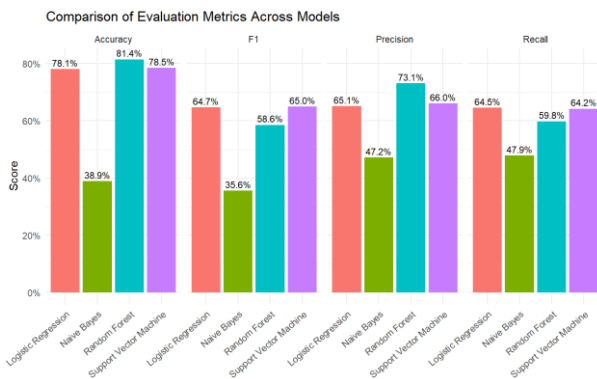


Figure 5 Comparison of Evaluation Metrics Across Models

IV. DISCUSSION

The Naïve Bayes classifier's noticeably weaker performance likely comes down to its conditional independence assumption, which doesn't capture the sparse, correlated nature of TF-IDF features well, and possibly a mismatch between the probability distribution the algorithm assumes and the continuous nature of TF-IDF weighted values (Kohli, 2025).

Although Random Forest had the highest overall accuracy, its very low recall for the Neutral class shows this result is mostly driven by the dominant Positive class. Put differently, the model is really just good at the majority class and struggles everywhere else so the impressive accuracy figure is masking a real problem with generalizing across all three sentiment categories. That lines up with what earlier work has already pointed out: ranking models by accuracy alone can be deceptive once the dataset is imbalanced, and macro-averaged metrics such as macro F1-score tend to give a more honest read on how well a model actually generalizes(Farhadpour et al., 2024).

SVM tells a different story it posted the best macro F1-score of the four, which suggests a more even, dependable performance across all three classes. That's probably down to how well SVM handles the high-dimensional, sparse feature space TF-IDF tends to produce(Al-Qudah et al., 2025). Logistic Regression wasn't far behind it on macro F1 either, staying stable and consistent as a baseline linear model something that's been observed with TF-IDF-based Logistic Regression in other review domains too(Zahra et al., 2026).

The comparison here points to a real gap between accuracy and macro F1-score as evaluation criteria on this imbalanced dataset. This gap shows why macro-averaged F1-score should be used alongside accuracy, rather than accuracy alone, when evaluating sentiment

classification models on datasets with skewed class distributions, in line with recent literature on imbalanced classification (Farhadpour et al., 2024). Based on macro F1-score, SVM comes out as the most suitable model for multiclass sentiment classification of restaurant reviews in this study, a conclusion that broadly matches recent restaurant-review sentiment studies favoring SVM or TF-IDF-based linear classifiers over Naïve Bayes for balanced, generalizable performance (Kohli, 2025).

V.CONCLUSION

This paper carried out a systematic comparison of four supervised machine learning algorithms Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine for multiclass sentiment classification of restaurant reviews. The dataset, made up of 9,948 valid reviews after cleaning, was labelled into three sentiment categories based on numerical ratings and run through a standardized pipeline of text preprocessing, TF-IDF feature extraction, and variance-based feature selection. Random Forest ended up with the best overall accuracy at 81.45%, but it barely caught any of the minority Neutral class a real limitation once you look past the headline number. SVM told a more balanced story: 78.53% accuracy paired with the best macro F1-score of the four (65.03%), spread more evenly across all three sentiment categories. Naïve Bayes, for its part, trailed behind on every metric. Taken together, this points to a simple conclusion: picking a model for imbalanced sentiment classification can't be based on accuracy alone macro averaged metrics that actually account for minority classes need to be part of the decision too. More broadly, how the text is preprocessed, how features are extracted via TF-IDF, and which metric gets used for evaluation all turn out to matter quite a bit when the dataset is real-world and imbalanced, as restaurant reviews typically are. There's room to push this further. Class-imbalance mitigation oversampling, class-weighted learning, that kind of thing could help with the minority-class problem specifically. Word embeddings like Word2Vec or FastText are another direction worth trying, along with ensemble boosting methods Cross validation and hyperparameter tuning would also be reasonable next steps for making the models more generalizable and robust.

REFERENSI

- Al-Qudah, D. A., Ala'M, A.-Z., Cristea, A. I., Merelo-Guervós, J. J., Castillo, P. A., & Faris, H. (2025). Prediction of sentiment polarity in restaurant reviews using an ordinal regression approach based on evolutionary XGBoost. *PeerJ Computer Science*, 11, e2370.
- Barron, R. C. (2025). *Trustworthy Domain-Specific AI for Structured Knowledge Retrieval and Reasoning*. University of Maryland, Baltimore County.
- Farhadpour, S., Warner, T. A., & Maxwell, A. E. (2024). Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote Sensing*, 16(3), 533.
- Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Overview of the Transformer-based Models for NLP Tasks. *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 179–183.
- Haiman, J. (1998). *Talk is cheap: Sarcasm, alienation, and the evolution of language*. Oxford University Press.
- Kamath, B. P., Geetha, M., Acharya, U. D., Singh, D., Rao, A., Rai, S., & Shetty, R. (2025). Comprehensive analysis of word embedding models and design of effective feature vector for classification of Amazon product reviews. *IEEE Access*, 13, 25239–25255.
- Kohli, K. (2025). *Sentiment Analysis of Restaurant Reviews Using Machine Learning Algorithms*.
- Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University Computer and Information Sciences*, 36(4), 102048.

- Mehedi, M. H. K., Farid, F. Al, Rhythm, E. R., Rahman, F., Hasib, K. M., Uddin, J., & Mansor, S. (2025). Deep hierarchical networks for sentiment analysis of restaurant reviews from food apps. *Scientific Reports*, 15(1), 40148.
- Pajila, P. J. B., Sheena, B. G., Gayathri, A., Aswini, J., & Nalini, M. (2023). A comprehensive survey on naive bayes algorithm: Advantages, limitations and applications. *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, 1228–1234.
- Pradana, A., Fadllullah, A., Prasetya, A., & Fadliansyah, F. (2025). The Application Of The Multi-Objective Optimization On The Basis Of Simple Ratio Analysis Method In A Decision Support System For Prospective Ubt Student Association Chair Candidates. *JIPPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, 10(2), 1804–1812.
- Rahman, A., Khan, M. A., Bishnu, K. K., Rozario, U., Ishraq, A., Mridha, M. F., & Aung, Z. (2025). Multilingual sentiment analysis in restaurant reviews using aspect focused learning. *Scientific Reports*, 15(1), 28371.
- Rokhva, S., Teimourpour, B., & Babaei, R. (2025). Enhanced sentiment analysis of Iranian restaurant reviews utilizing sentiment intensity analyzer & fuzzy logic. *Food and Humanity*, 5, 100658.
- Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7), 5731–5780.
- Yarsasi, S., & Iskoko, A. (2025). A Text Classification Approach for Detecting Cyberbullying Risk on Twitter Using Support Vector Machine with Naive Bayes and Random Forest Comparison. *International Journal of Informatics and Information Systems*, 8(4), 175–187.
- Zahra, N. Z., Farhanatussaidah, S., Afifah, N. N., Muthoharoh, L., Satria, A., & Manullang, M. C. T. (2026). Benchmarking Logistic Regression, SVM, Naive Bayes, and IndoBERT fine-tuning for sentiment analysis on Indonesian product reviews. *ArXiv Preprint ArXiv:2605.03439*.