Edu Komputika 12 (1) (2025)

Edu Komputika Journal

https://journal.unnes.ac.id/journals/edukom/

Weakly Supervised Sentiment Analysis of Indonesian Rural Tourism Reviews: A TF-IDF Baseline for Melung Tourism Village

Zanuar Rifa'i[1)] and Bayu Priya Mukti[2)]✉

[1]Department of Digital Business, Faculty of Business and Social Science, Amikom Purwokerto University, Indonesia

[2]Master of Computer Science Study Program, Faculty of Computer Science, Amikom Purwokerto University, Indonesia

Article Info

Article History:
Received: 6 August 2025
Revised: 21 December 2025
Accepted: 23 December 2025

Keywords:
Indonesian Tourist Reviews, Machine Learning Classification, Random Oversampling, Sentiment Analysis, TF-IDF Feature Extraction

Abstract

This study investigates sentiment classification of Indonesian-language tourist reviews from the rural destination of Melung Tourism Village. A total of 724 user-generated reviews from 546 unique users are preprocessed using Indonesian-specific text cleaning, stopword filtering, and stemming, then weakly labeled through a stemmed positive–negative lexicon. TF-IDF unigram–bigram features are extracted from the preprocessed texts and used to train three classical classifiers: Naive Bayes, linear Support Vector Machine (SVM), and Logistic Regression. To address class imbalance, RandomOverSampler is applied only to the training data, and model evaluation combines stratified 5-fold cross-validation with a held-out test set, using weighted F1-score as the primary metric. Logistic Regression achieves the best performance on the test set (weighted F1 = 0.8799, accuracy = 0.8828), closely followed by SVM, while Naive Bayes lags behind. The results show that, even with a modest, weakly supervised dataset, a carefully designed classical pipeline can yield reliable sentiment indicators to support data-driven management of rural tourism destinations.

✉ Correspondence Address:
Jl. Letjend Pol. Soemarto No.127, Watumas, Purwanegara, Kec. Purwokerto Utara, Kabupaten Banyumas, Jawa Tengah 53127
E-mail: bayupriyaa@amikompurwokerto.ac.id

ISSN 2252-6811
E-ISSN 2599-297X

INTRODUCTION

The advancement of digital technology has significantly transformed the way people access and share information, including in the tourism sector. One of the most prevalent forms of user-generated content is online reviews or testimonials on products and services, which have become an important element in shaping public perception (Osly Usman & Wijaya, 2025; Saini & Mishra, 2025). In the tourism context, reviews submitted by travelers on digital platforms not only reflect their personal experiences but also serve as valuable references for prospective tourists in making travel decisions. Consequently, review data has emerged as a critical source of insight that can be further explored for data-driven decision-making by tourism destination managers.

Sentiment analysis has become one of the primary approaches in Natural Language Processing (NLP) to extract emotional information from unstructured textual data. This technique enables the categorization of opinions into sentiment classes such as positive, neutral, and negative, thereby facilitating systematic interpretation of user feedback. In the tourism domain, the ability to automatically classify sentiment in tourist reviews presents strategic opportunities for destination managers to evaluate visitor satisfaction, identify service weaknesses, and adjust promotional strategies in real-time (Sreenivas et al., 2023). Such capabilities contribute directly to enhancing destination competitiveness and promoting sustainable tourism management based on visitor experiences.

However, applying sentiment analysis to Indonesian-language texts—particularly in a three-class classification setting—presents unique linguistic and computational challenges. The Indonesian language exhibits rich morphology, informal expressions, and frequent code-mixing, especially in user-generated content. Moreover, imbalanced sentiment distributions, where positive reviews dominate, further complicate model training. Addressing these challenges requires careful preprocessing, feature extraction, and balancing strategies.

Previous studies have proposed a range of methods to improve sentiment analysis performance. Commonly, Term Frequency–Inverse Document Frequency (TF-IDF) has been adopted for text representation due to its interpretability and efficiency. Although modern embedding techniques such as Word2Vec or BERT offer semantic richness, TF-IDF remains well-suited for small- to medium-scale datasets by providing stable, transparent, and computationally efficient representations (Ondara et al., 2022). Given the relatively small dataset, TF-IDF was selected to maintain model interpretability and prevent overfitting, which can occur when using dense embeddings on limited data (Choi & Lee, 2017). Furthermore, data balancing techniques such as Random Oversampling have also been utilized to improve model performance when dealing with imbalanced class distributions (Bhattacharjee et al., 2021). While more sophisticated techniques like SMOTE could also be explored, Random Oversampling was chosen because it avoids generating synthetic text vectors that might distort semantic integrity—an important consideration for short, informal Indonesian sentences. This approach prioritizes linguistic fidelity over algorithmic complexity, consistent with the exploratory nature of this study (Deniz et al., 2021).

Equally important is the selection of classification algorithms, which play an important role in determining the success of a sentiment analysis system. Algorithms such as Naive Bayes, Support Vector Machine (SVM), and Logistic Regression have been widely applied and compared in text analysis studies (Sahu & Selot, 2022). Naive Bayes is known for its computational efficiency in handling large-scale textual data under the assumption of feature independence (Kiran Kumar, Prajwal, & Nivedita, 2024). SVM is effective in cases with clear margins between classes, while Logistic Regression excels in its interpretability. However, the performance of these algorithms is highly context-dependent, particularly influenced by the data characteristics and preprocessing strategies employed. Therefore, comparative studies of algorithm performance in specific contexts—such as Indonesian-language rural tourism reviews—remain highly relevant and necessary for deriving actionable insights.

Several studies have adopted similar sentiment analysis approaches within the tourism sector. For instance, lexicon-based methods using positive and negative word lists can be effectively applied for automatic sentiment labeling in government social media data (Aksu & Karaman, 2021)(Saraswati et al., 2024). Meanwhile, oversampling techniques could enhance sentiment classification accuracy in the context of local Indonesian tourism reviews (Fatah et al., 2024). Nevertheless, there remains a lack of research explicitly comparing the effectiveness of different classification algorithms in the specific setting of rural Indonesian tourism, especially with consideration of preprocessing variations and class balancing strategies.

To address this research gap, the present study develops a comparative approach involving

three classification algorithms—Naive Bayes, Support Vector Machine, and Logistic Regression—for sentiment analysis of tourist reviews from Desa Wisata Melung. The study focuses on evaluating the performance of each algorithm in classifying Indonesian-language reviews into three sentiment categories (positive, neutral, negative) under realistic data and resource constraints. Methodologically, the proposed pipeline first applies Indonesian-specific text preprocessing—including lowercasing, noise removal, stopword filtering, and stemming—followed by weakly supervised lexicon-based sentiment labeling, training-set balancing with RandomOverSampler, and TF-IDF unigram–bigram feature extraction. These features are then used to train and compare Naive Bayes, SVM, and Logistic Regression models using stratified 5-fold cross-validation and a held-out test set, with weighted F1-score as the primary evaluation metric and macro F1 as a complementary measure. Ultimately, this research is expected to identify an effective yet computationally affordable technical approach and contribute practical value in developing automated sentiment monitoring systems for tourism destination management, particularly in rural contexts.

RESEARCH METHODS

This section systematically outlines the methodological steps undertaken to develop and evaluate a machine learning-based sentiment classification model for Indonesian-language tourism reviews (Panjaitan, 2025). The methodology is designed to ensure replicability and empirical validity, encompassing data collection and cleaning, text transformation into numerical features, class distribution balancing, automated sentiment labeling using a lexicon-based approach, model selection and training, as well as performance evaluation using relevant classification metrics. The overall research workflow is illustrated in Figure 1.

A. Collecting Data

This study utilized secondary textual data in the form of user-generated reviews from visitors to the Melung Tourism Village, obtained from public platforms such as Google Maps, Instagram, and TikTok. These platforms were selected due to their accessibility and popularity among domestic tourists, ensuring diversity of content and opinion.

The scraping process gathered 724 reviews from 546 unique users, providing sufficient linguistic variation for exploratory sentiment classification. Although this sample size is modest, previous studies have shown that sentiment classification performance is highly sensitive to dataset size, yet small corpora remain valid for exploratory and comparative analyses when interpretability is prioritized over generalization (Choi & Lee, 2017).

An initial validation step was conducted to ensure data completeness, and any entries containing missing values in the review column were removed to maintain the integrity and quality of the dataset (Lubihana & Y., 2022).

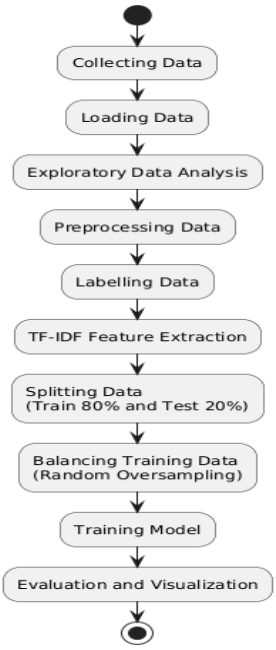

Figure 1. Research flow

B. Data Preprocessing

Prior to analysis, the dataset was thoroughly examined to ensure completeness and validity. Rows containing missing entries, particularly in the review text column, were removed to avoid distortion in subsequent processing and sentiment labeling. Once filtered, the data underwent a text preprocessing phase to convert raw textual input into a more structured format suitable for machine learning analysis. The preprocessing procedures included converting all characters to lowercase for input standardization, removing non-alphabetic characters such as numbers, symbols, and punctuation marks, and eliminating redundant whitespace. Unlike many sentiment analysis workflows, this study intentionally excluded stopword removal and stemming. Indonesian stopwords often carry syntactic or semantic importance, and stemming can distort contextual meaning—particularly in phrases like "*tidak terlalu ramai*" (not too crowded), where removing function words alters sentiment polarity.

Maintaining these words helps preserve subtle emotional and contextual cues (Duong & Nguyen-Thi, 2021)(Pradana & Hayaty, 2019). Recent studies have also shown that aggressive preprocessing can reduce performance in morphologically rich or low-resource languages (Shehu et al., 2021). Therefore, the decision to retain linguistic richness was made to prioritize semantic accuracy over feature sparsity.

The final step was tokenization, where each sentence was segmented into individual word units (tokens). These processed tokens were then used as input for further analysis using machine learning models.

C. Exploratory Data Analysis

Following text preprocessing, an initial exploratory data analysis (EDA) was conducted on the review dataset. This analysis examined the distribution of character lengths and word counts per review to assess the density and informativeness of each entry. In addition, two complementary word clouds were generated to visualize frequently mentioned terms.

The first word cloud is based on minimally processed text (lowercasing and basic cleaning while retaining stopwords), illustrating how raw user-generated content is dominated by function words and informal particles. The second word cloud is built from the enhanced, Indonesian-specific preprocessing pipeline (stopword filtering and stemming), which suppresses non-informative function words and highlights content-bearing terms such as "kolam", "renang", "wisata", "sejuk", and "sawah". These word clouds are used purely as qualitative exploratory visualizations to reveal dominant themes in the corpus and are not used as features for the classification models (Zhu et al., 2024) (Da Poian et al., 2023). This process served to validate whether the collected data accurately reflected relevant tourism experiences and also provided a preliminary basis for observing potential sentiment distributions.

D. Sentiment Labelling

Sentiment labeling was carried out using a predefined lexicon of positive and negative words. This process followed a three-stage approach: first, texts containing no sentiment-related words were automatically labeled as neutral; second, texts with a minimal difference between the number of positive and negative words were also assigned a neutral label; and third, texts dominated by either positive or negative words were labeled accordingly. While manual annotation or deep-learning-based labeling could yield finer accuracy, such methods require extensive resources and are rarely feasible for low-resource languages like Indonesian. The lexicon-based method remains an efficient and transparent baseline in sentiment analysis, particularly when annotated datasets are limited (Touahri, 2022). Moreover, lexicon expansion techniques (e.g., Word2Vec-based enrichment) have been shown to improve coverage while maintaining interpretability (Alshari et al., 2018). This study employed a manually validated lexicon drawn from tourism-related Indonesian corpora to balance feasibility with contextual precision.

The positive and negative word lists used in this study were constructed from commonly occurring terms in Indonesian-language tourism reviews and are summarized in Table 1. To ensure full consistency with the preprocessing pipeline, all lexicon entries were subjected to the same normalization and stemming steps as the review texts: words were lowercased, stripped of punctuation and non-alphabetic characters, filtered for stopwords, and then stemmed using the Sastrawi-based Indonesian stemmer. During implementation, matching is performed on these stemmed forms (e.g., menyenangkan → senang, mengecewakan → kecewa), whereas Table 1 reports the corresponding surface forms for readability.

Each preprocessed review was then analyzed by counting the number of occurrences of positive and negative lexicon items. Reviews that did not contain any words from either category were labeled as neutral. If the number of positive words exceeded the number of negative words by at least one, the review was labeled as positive; conversely, if negative words were more dominant by the same threshold, the review was labeled as negative. Cases where the difference between positive and negative counts did not meet this threshold were assigned a neutral label. This procedure acts as a form of weak supervision in a low-resource setting: it enables scalable labeling of the corpus in the absence of a manually annotated gold standard, while remaining transparent and easy to replicate. At the same time, its limitations and the need for future work with human-annotated data and inter-annotator agreement analysis are explicitly acknowledged in the conclusion

Table 1. Examples of Positive and Negative Sentiment Words Used in the Lexicon (Surface Forms Before Stemming)

No.	Category	List of Words
1.	Positive	*bagus, baik, indah, sejuk, nyaman, mantap, puas, recommended, asri, adem, menyenangkan, murah, ramah, bersih, keren, cantik, seru, luas, jernih, betah, enak, menarik, terbaik, mewah, luar biasa, wow, hebat, top*
2.	Negative	*kotor, kurang, mahal, jauh, jelek, buruk, panas, kecil, sempit, rusak, ramai, capek, lelah, macet, bising, tidak, nggak, ngga, enggak, kapok, mengecewakan, membosankan, payah, salah, tidak nyaman, parah*

E. Data Balancing

Following the sentiment labeling process, it was found that the distribution of data across sentiment categories was imbalanced, with a predominance of positive reviews. This imbalance can introduce bias into the classification model, making it more likely to recognize patterns from the majority class while underperforming on minority classes (Wang et al., 2021). The original dataset was first split into training (80%) and test (20%) sets using a stratified sampling strategy to preserve the proportion of sentiment classes. In the training set, the distribution of 291 positive, 236 neutral, and 52 negative reviews was transformed into a perfectly balanced set of 291 instances per class by randomly duplicating minority-class examples using RandomOverSampler. The test set (73 positive, 59 neutral, 13 negative) remained untouched to provide an unbiased estimate of generalization.

Although synthetic data generation techniques such as SMOTE are popular, they can distort semantic relationships in textual data, as interpolated TF-IDF vectors may not correspond to linguistically valid expressions. Therefore, Random Oversampling was preferred to maintain the authenticity and interpretability of textual content (Deniz et al., 2021). In the cross-validation experiments, oversampling was likewise applied only to the training portion of each fold to avoid data leakage into validation splits.

F. Fitur Extraction Using TF-IDF

Text feature representation was carried out using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF assigns weights to words based on their frequency in an individual document relative to their occurrence across the entire corpus. This technique is effective in highlighting terms that carry specific meaning within a particular review, without overemphasizing commonly used words (Ningsih & Unjung, 2025) (Das et al., 2023). TF-IDF calculates the weight of words based on their frequency within a document relative to the corpus, emphasizing unique and informative terms. Both unigrams and bigrams were used to capture contextual phrases such as "tidak bagus" (not good) and "sangat indah" (very beautiful). Despite the rise of semantic embeddings like Word2Vec or BERT, TF-IDF remains a robust choice for small- to medium-scale datasets because of its interpretability, computational efficiency, and lower risk of overfitting (Ondara et al., 2022). Moreover, prior research demonstrates that TF-IDF can perform competitively with embeddings when combined with feature selection and balancing (Deniz et al., 2021).

G. Splitting Data

The dataset was then divided into two main subsets: training data and testing data, with a split ratio of 80 percent and 20 percent, respectively. The partitioning was performed randomly while preserving the proportional distribution of sentiment classes through a stratified sampling technique. This approach was intended to ensure that the model could effectively learn from the training data and be subsequently evaluated on previously unseen data to assess its generalization performance.

H. Training Model

The modeling process employed three widely used machine learning algorithms for text classification: Naive Bayes, Support Vector Machine (SVM), and Logistic Regression. These algorithms were selected because they are well-established baselines for document classification, handle high-dimensional sparse feature spaces efficiently, and can be trained and deployed with modest computational resources—an important consideration for village-level tourism managers. Their use is further supported by prior studies that highlight their proven reliability and computational simplicity in low-resource and small-data settings (Devi & Saharia, 2020). Although deep learning methods, such as LSTM and BERT, often achieve higher accuracy, they require substantially larger datasets and training resources. In contrast, conventional algorithms

can still perform competitively on smaller corpora when preprocessing and feature engineering are carefully optimized (Romadhony et al., 2024). In this study, each model was trained on preprocessed and TF-IDF–represented reviews. The original dataset was first split into training (80%) and test (20%) sets using a stratified sampling strategy to preserve the proportion of sentiment classes. To address class imbalance, RandomOverSampler was applied only to the training set; the test set remained untouched to provide an unbiased estimate of generalization. Specifically, the training distribution of 291 positive, 236 neutral, and 52 negative reviews was transformed into a perfectly balanced set of 291 instances per class, while the test set preserved its original distribution of 73 positive, 59 neutral, and 13 negative reviews. In addition to this hold-out evaluation, each classifier was also embedded in an imbalanced-learning pipeline combining RandomOverSampler and the classifier within a Stratified 5-Fold Cross Validation (Stratified K-Fold) framework. In every fold, oversampling was applied only to the training portion of that fold, thereby avoiding data leakage into validation splits and allowing a more robust estimate of model performance and stability across different data partitions. SVM and Logistic Regression were trained with a linear kernel / linear decision boundary and class-weight="balanced" option to compensate for residual imbalance in the original labels, while Naive Bayes (MultinomialNB) operated directly on TF-IDF counts; all models were implemented using scikit-learn with default hyperparameters. Model evaluation employed four standard metrics: accuracy, precision, recall, and F1-score. Due to the initial class imbalance, F1-score was prioritized as the primary performance indicator, since it balances precision and recall and provides a fairer assessment for minority classes (Burns et al., 2011).

I. Evaluation and Visualization

Model performance evaluation was conducted using several standard classification metrics, including accuracy, precision, recall, and F1-score. Additionally, confusion matrices were provided to illustrate the number of correct and incorrect predictions for each sentiment class. These metrics were particularly important given the initially imbalanced dataset, where accuracy alone would be insufficient to reflect the model's true performance. The F1-score was used to present a harmonic mean between precision and recall, which is especially relevant for assessing the model's effectiveness in identifying minority classes.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad (2)$$

$$\text{F1 Score} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \quad (3)$$

In addition to scalar metrics, confusion matrices were generated to illustrate the number of correct and incorrect predictions for each sentiment class, enabling a more fine-grained interpretation of model behavior and error patterns. For the cross-validation experiments, the mean and standard deviation of accuracy, weighted F1, and macro F1 across the five folds were computed and visualized using boxplots to assess score stability. Finally, cross-validation and test-set performances were compared side by side to detect potential overfitting or optimistic bias.

All modeling and evaluation procedures were implemented in the Python programming environment, utilizing supporting libraries such as scikit-learn, pandas, and Indonesian-language NLP toolkits. The hardware setup used for the experiments included an Intel Core i5 processor with 24GB RAM, which was adequate for processing and training classification models on this medium-scale dataset.

RESULT AND DISCUSSION

An initial analysis of the review dataset for the Melung tourist destination was conducted to gain a comprehensive understanding of the data structure and characteristics used in the sentiment classification process. The dataset consisted of a total of 724 reviews submitted by 546 unique users. This indicates that most users contributed only one review, reflecting a relatively high diversity of opinions. A detailed summary of the dataset is presented in Table 2.

The initial sentiment distribution across the three categories revealed a significant imbalance, with a predominance of positive reviews (364) and neutral ones (295), while negative reviews were in the minority (65). After the 80/20 stratified split, the training set comprised 291 positive, 236 neutral, and 52 negative reviews, while the test set contained 73 positive, 59 neutral, and 13 negative reviews. RandomOverSampler was then applied only to the training set to produce a balanced training distribution of 291 instances per class, leaving the test set unchanged. This class disparity posed a

risk of prediction bias in machine learning models, as models tend to favor the majority class, and therefore justified the use of both balancing strategies and F1-based evaluation metrics.

Table 2. Statistics Data of Melung Tourism Reviews

Matrices	Value
Total Number of Reviews	724
Number of Unique Reviewers	546
Missing Values (reviewer_name)	3
Missing Values (review_text)	0
Distribution of Reviews Sentimen	Positive: 364, Neutral: 295, Negative: 65

Exploratory visual analysis using the paired word clouds in Figure 2 provides further insight into the lexical content of the reviews. In the left panel, which is constructed from minimally processed text with stopwords retained, high-frequency function words such as "yang", "dan", "untuk", and "juga" dominate the visualization, masking some of the underlying thematic content. In contrast, the right panel, which is generated after the enhanced preprocessing stage (stopword removal and stemming), reveals that terms such as "kolam" (pool), "renang" (swimming), "wisata" (tourism), "sejuk" (cool), "alam" (nature), and "sawah" (rice field) appear most prominently. This indicates that the physical and environmental attributes of the destination are a primary focus for visitors, confirming the strong contextual alignment of the dataset with rural nature-based tourism.

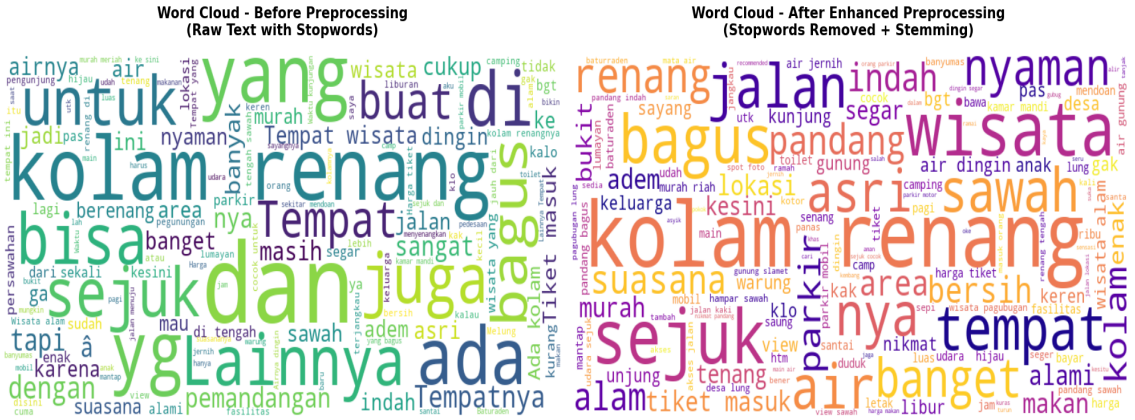

Figure 2. Word clouds of visitor reviews: before enhanced preprocessing (raw text with stopwords retained), and after enhanced preprocessing (stopwords removed and stemming applied).

To identify sentiment categories within the reviews, three machine learning algorithms—Naive Bayes, Support Vector Machine (SVM), and Logistic Regression—were trained and compared. Evaluation was conducted using four primary metrics: accuracy, precision, recall, and F1-score, across the three sentiment classes (positive, neutral, and negative).

Given the original class imbalance, accuracy alone is not a reliable indicator of model performance, as it can overrepresent the success of the majority class. Instead, the F1-score was emphasized as the primary metric because it balances precision and recall, providing a more robust measure of performance for minority classes (Burns et al., 2011). Table 3 presents the performance of the three classifiers on the held-out test set, using accuracy, weighted F1-score, macro F1-score, precision, and recall.

Table 3. Test-Set Performance with Weighted F1 as Main Metric

Models	Accuracy	Wighted F1	Macro F1	Precision	Recall
Naive Bayes	0.7241	0.7238	0.6424	0.7785	0.7241
SVM (linear)	0.8759	0.8745	0.8307	0.8744	0.8759
Logistic Regression	0.8828	0.8799	0.8283	0.8805	0.8828

The results show that Logistic Regression achieved the highest weighted F1-score (0.8799) and accuracy (0.8828), closely followed by SVM (weighted F1 = 0.8745; accuracy = 0.8759). Naive Bayes lagged noticeably behind, with a weighted F1-score of 0.7238. The macro F1-scores—0.6424 for Naive Bayes, 0.8307 for SVM, and 0.8283 for Logistic Regression—indicate that both SVM and Logistic Regression handle all three classes more evenly than Naive Bayes, which struggles particularly with the positive and neutral categories. Because the dataset is originally imbalanced, the emphasis on weighted F1-score rather than raw accuracy directly addresses concerns about evaluation bias. Weighted F1 captures a trade-off between precision and recall while accounting for the number of test instances in each class, making it a more informative indicator of model performance on this task.

A more detailed view of model behaviour is obtained by examining per-class precision, recall, and F1-scores alongside the confusion matrices (Figure 3). On the test set: (1) For the negative class, both Logistic Regression and SVM achieve high precision and recall, with F1-scores above 0.92, indicating that strongly negative reviews are captured reliably, despite being the minority; (2) For the neutral class, SVM and Logistic Regression again perform comparably with F1-scores in the mid-0.8 range, whereas Naive Bayes shows substantially lower performance; and (3) For the positive class, which is relatively small in the test set, SVM and Logistic Regression obtain F1-scores around 0.70, while Naive Bayes trails behind, reflecting difficulties in separating mildly positive content from neutral comments.

Overall, the confusion matrices show that most errors occur between positive and neutral reviews, which is intuitive given that many tourism comments express mild appreciation, making the boundary between "slightly positive" and "neutral" inherently fuzzy. Misclassification between positive/neutral and negative is less frequent, suggesting that the lexicon-based labeling and TF-IDF features capture strongly polarized expressions reasonably well. Figure 4 shows Stratified 5-fold cross-validation boxplots for accuracy, weighted F1, macro F1, and precision for the three classifiers.

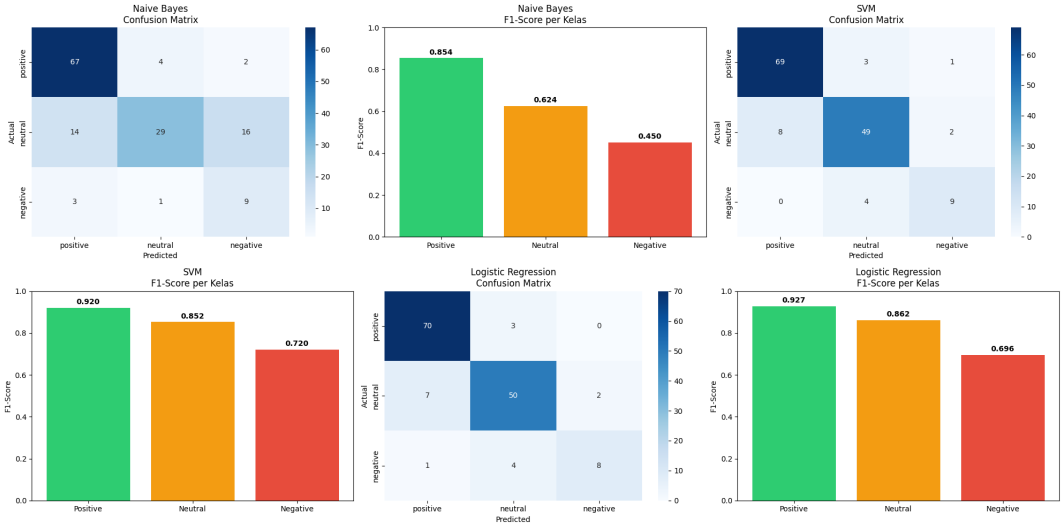

Figure 3. Confusion matrices and per-class F1-scores for Naive Bayes, SVM, and Logistic Regression on the test set.

Table 4. Mean Stratified 5-Fold Cross-Validation Performance

Models	CV Accuracy (mean ± sd)	CV F1-Weighted (mean ± sd)	CV Macro F1 (mean ± sd)
Naive Bayes	0.7617 ± 0.0626	0.7669 ± 0.0584	0.6805 ± 0.0740
SVM (linear)	0.8205 ± 0.0527	0.8145 ± 0.0561	0.7311 ± 0.0825
Logistic Regression	0.8326 ± 0.0685	0.8267 ± 0.0697	0.7415 ± 0.0868

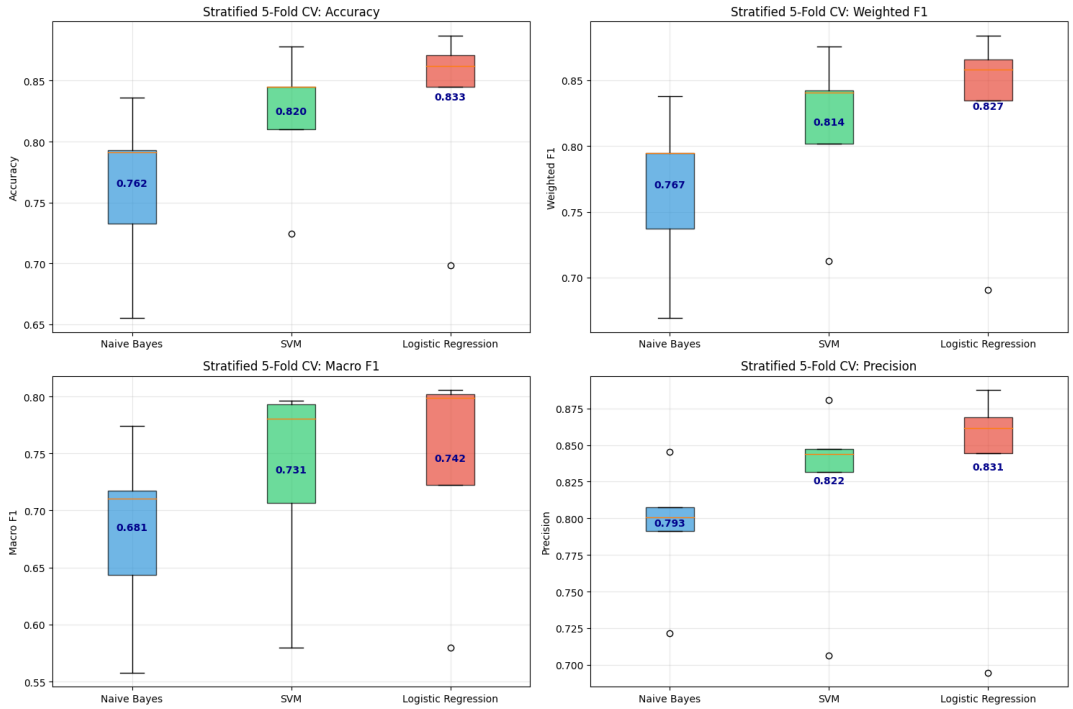

Figure 4. Stratified 5-fold cross-validation boxplots for accuracy, weighted F1, macro F1, and precision for the three classifiers

Table 5 compares the cross-validation weighted F1 with the test-set weighted F1 for each model, along with the absolute difference between them.

Table 5. Mean Stratified 5-Fold Cross-Validation Performance

Models	CV F1-Weighted	Test F1-Weighted	CV-Test
Logistic Regression	0.8267	0.8799	0.0532
SVM (linear)	0.8145	0.8745	0.0600
Naïve Bayes	0.7669	0.7238	0.0431

For all three models, the difference between cross-validation and test weighted F1-scores is in the range of 0.04–0.06. This pattern suggests that, although the test-set performance of Logistic Regression and SVM is somewhat higher than their mean cross-validation performance, there is no evidence of extreme overfitting: the models generalize reasonably well across folds and to the held-out test set.

Figure 5 further illustrates this relationship by plotting accuracy and weighted F1 for both cross-validation and test evaluations. For each model, the test bars are slightly higher than the cross-validation bars, but the differences remain moderate.

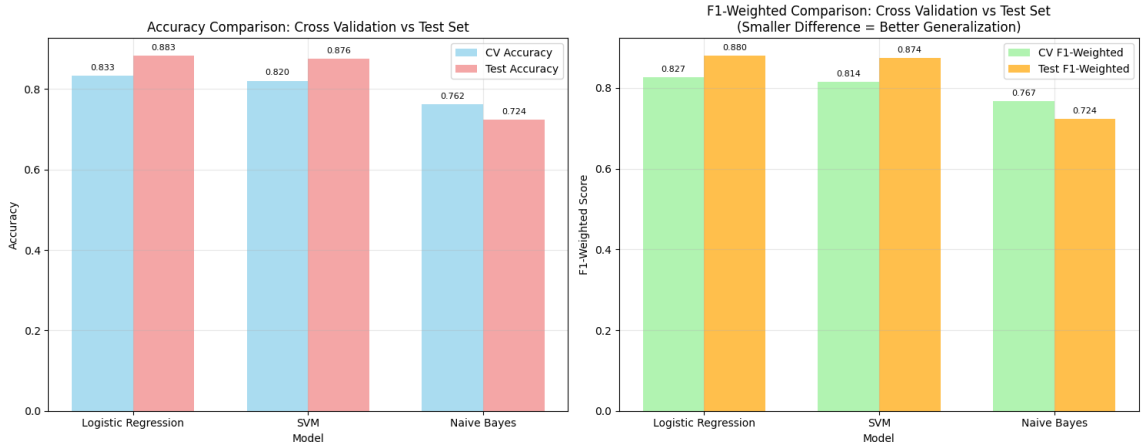

Figure 5. Comparison of cross-validation and test-set performance (accuracy and weighted F1) for Naive Bayes, SVM, and Logistic Regression

From a substantive standpoint, the high F1-scores of SVM and Logistic Regression—particularly for the negative class—indicate that the models can reliably detect dissatisfied visitors even when such reviews are relatively rare. This capability is crucial for rural tourism managers who must respond promptly to critical feedback about facilities, access, or services. When combined with the qualitative insights from exploratory analysis (e.g., frequent terms related to kolam, renang, sawah, and sejuk), the sentiment classification results provide a quantitative layer that complements traditional monitoring of tourism experiences. The pipeline outlined in this study thus offers a practical tool for converting a modest volume of online reviews into actionable sentiment indicators, even in resource-constrained settings.

CONCLUSION

This study evaluated and compared the performance of three machine learning algorithms—Naive Bayes, Support Vector Machine (SVM), and Logistic Regression—for multiclass sentiment classification (positive, neutral, negative) of tourist reviews about the Melung Tourism Village. Using Indonesian-specific preprocessing (including stopword removal and stemming), weakly supervised lexicon-based labeling, TF-IDF unigram–bigram features, Random Oversampling on the training set, stratified 5-fold cross-validation, and a held-out test set, SVM and Logistic Regression consistently outperformed Naive Bayes. On the test set, Logistic Regression achieved the highest weighted F1-score (0.8799) and accuracy (0.8828), with SVM very close behind (weighted F1 = 0.8745), while Naive Bayes lagged substantially.

From a methodological perspective, the study demonstrates that weighted F1-score—rather than accuracy alone—provides a more appropriate evaluation metric for imbalanced sentiment data. The cross-validation analysis further shows that performance is relatively stable across folds, and the differences between cross-validation and test weighted F1-scores remain moderate (approximately 0.04–0.06), suggesting that the best-performing models do not severely overfit the data.

Beyond these quantitative findings, the study offers a complete and transparent pipeline for Indonesian-language sentiment analysis in a realistic rural tourism context: from multichannel data collection and weakly supervised lexicon-based labeling to balancing strategies, TF-IDF feature extraction, and comparative evaluation of classical machine learning algorithms under cross-validation. This pipeline is intentionally designed to be reproducible and computationally affordable for practitioners.

From an applied standpoint, the results show that even with a relatively small number of reviews (724 from 546 unique users), it is possible to extract reliable sentiment indicators that reflect visitor perceptions of a rural tourism destination. Positive sentiment dominates, and recurring terms such as "sejuk," "alami," and "sawah" reinforce Melung's image as a nature-based village destination. The ability of the models to accurately flag negative reviews is particularly valuable for prioritizing service improvements and managing reputation.

Several limitations remain. First, the dataset is restricted to a single rural destination, limiting generalizability to other types of destinations. Second, sentiment labels were obtained through a lexicon-based weak

supervision scheme without manual annotation or inter-annotator agreement analysis, which may overlook nuanced linguistic phenomena. Third, the balancing strategy relies on Random Oversampling, which, while simple and interpretable, could still magnify noise in rare classes. Fourth, although this study incorporates Stratified 5-Fold Cross Validation, evaluation is still confined to a single dataset; an external test set from other destinations would be required for a stronger generalization claim. Fifth, feature representation is limited to TF-IDF, and the model comparison focuses on classical algorithms (Naive Bayes, SVM, Logistic Regression). Deep learning models and semantic embeddings such as Word2Vec or Indonesian BERT, have not yet been explored.

Future research should (i) extend the corpus to multiple destinations and larger review sets, (ii) construct a manually annotated gold-standard dataset with inter-annotator agreement analysis, (iii) investigate more advanced resampling or cost-sensitive methods for imbalanced text classification, and (iv) benchmark classical baselines against deep learning architectures and embedding-based representations. Incorporating aspect-based sentiment analysis and explainable AI techniques would also enable more fine-grained insights into specific dimensions of tourism experiences.

Despite these limitations, the present work provides a contextual and methodological baseline for sentiment analysis in Indonesian rural tourism and shows that, under realistic data and resource constraints, carefully designed classical models can already yield useful and interpretable sentiment indicators.

ACKNOWLEDGEMENTS

This research was supported by the Institute for Research and Community Service (LPPM) of Universitas Amikom Purwokerto through the Research Grant Scheme for Strengthening Research Centers (Skema Hibah Penelitian Penguatan Pusat Studi). The authors express their sincere gratitude for the material and non-material support provided, which significantly contributed to the successful completion of this study.

REFERENCES

Aksu, M. Ç., & Karaman, E. (2021). Analysis of Turkish Sentiment Expressions About Touristic Sites Using Machine Learning. Journal of Intelligent Systems: Theory and Applications, 4(2), 103–112. https://doi.org/10.38016/jista.854250

Alshari, E. M., Azman, A., Doraisamy, S., Mustapha, N., & Alkeshr, M. (2018). Effective Method for Sentiment Lexical Dictionary Enrichment Based on Word2Vec for Sentiment Analysis. 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), 1–5. https://doi.org/10.1109/INFRKM.2018.8464775

Bhattacharjee, M., Ghosh, K., Banerjee, A., & Chatterjee, S. (2021). Multilabel Sentiment Prediction by Addressing Imbalanced Class Problem Using Oversampling. In S. Banerjee & J. K. Mandal (Eds.), Advances in Smart Communication Technology and Information Processing (Vol. 165, pp. 239–249). Springer Singapore. https://doi.org/10.1007/978-981-15-9433-5_23

Burns, N., Bi, Y., Wang, H., & Anderson, T. (2011). Sentiment Analysis of Customer Reviews: Balanced versus Unbalanced Datasets. In A. König, A. Dengel, K. Hinkelmann, K. Kise, R. J. Howlett, & L. C. Jain (Eds.), Knowledge-Based and Intelligent Information and Engineering Systems (Vol. 6881, pp. 161–170). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23851-2_17

Choi, Y., & Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. Information Systems Frontiers, 19(5), 993–1012. https://doi.org/10.1007/s10796-017-9741-7

Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. Frontiers in Astronomy and Space Sciences, 10, 1134141. https://doi.org/10.3389/fspas.2023.1134141

Das, M., K., S., & Alphonse, P. J. A. (2023). A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2308.04037

Deniz, A., Angin, M., & Angin, P. (2021). Evolutionary Multiobjective Feature Selection for Sentiment Analysis. IEEE Access, 9, 142982–142996. https://doi.org/10.1109/ACCESS.2021.3118961

Devi, M. D., & Saharia, N. (2020). Learning Adaptable Approach to Classify Sentiment with Incremental Datasets. Procedia Computer Science, 171, 2426–2434. https://doi.org/10.1016/j.procs.2020.04.262

Duong, H.-T., & Nguyen-Thi, T.-A. (2021). A review: Preprocessing techniques and data augmentation for sentiment analysis. Computational Social Networks, 8(1), 1. https://doi.org/10.1186/s40649-020-00080-x

Fatah, D. A., Rochman, E. M. S., Setiawan, W., Aulia, A. R., Kamil, F. I., & Su'ud, A. (2024). Sentiment Analysis of Public Opinion Towards Tourism in Bangkalan Regency Using Naïve Bayes Method. E3S Web of Conferences, 499, 01016. https://doi.org/10.1051/e3sconf/202449901016

Lubihana, E., & Y., B. (2022). Design of a Tourism Recommendation System Based on Sentiment Analysis with Lexicon LSTM. 2022 International Symposium on Electronics and Smart Devices (ISESD), 1–6. https://doi.org/10.1109/ISESD56103.2022.9980738

Moreo, A., Esuli, A., & Sebastiani, F. (2016). Distributional Random Oversampling for Imbalanced Text Classification. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 805–808. https://doi.org/10.1145/2911451.2914722

Ningsih, M. R., & Unjung, J. (2025). Sentiment Analysis on SocialMedia Using TF-IDF Vectorization and H2O Gradient Boosting for Student Anxiety Detection. Scientific Journal of Informatics, 11(4), 1137–1144. https://doi.org/10.15294/sji.v12i1.20582

Ondara, B., Waithaka, S., Kandiri, J., & Muchemi, L. (2022). Machine Learning Techniques, Features, Datasets, and Algorithm Performance Parameters for Sentiment Analysis: A Systematic Review. Open Journal for Information Technology, 5(1), 1–16. https://doi.org/10.32591/coas.ojit.0501.01001o

Osly Usman, & Wijaya, C. N. S. (2025). The Influence of Social Proof and User-Generated Content (UGC) on Brand Perception through Consumer Trust among Digital Consumers. International Student Conference on Business, Education, Economics, Accounting, and Management (ISC-BEAM), 3(1), 2654–2673. https://doi.org/10.21009/ISC-BEAM.013.191

Panjaitan, C. H. P. (2025). Systematic Literature Review of Sentiment Analysis on Various Review Platforms in the Tourism Sector. Journal of Advanced Computer Knowledge and Algorithms, 2(1), 12–18. https://doi.org/10.29103/jacka.v2i1.20287

Pradana, A. W., & Hayaty, M. (2019). The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 375–380. https://doi.org/10.22219/kinetik.v4i4.912

Romadhony, A., Al Faraby, S., Rismala, R., Wisesty, U. N., & Arifianto, A. (2024). Sentiment Analysis on a Large Indonesian Product Review Dataset. Journal of Information Systems Engineering and Business Intelligence, 10(1), 167–178. https://doi.org/10.20473/jisebi.10.1.167-178

Sahu, M. K., & Selot, S. (2022). Comparative Analysis of Various Supervised Machine Learning Techniques Used for Sentiment Analysis on Tourism Reviews. In R. P. Mahapatra, S. K. Peddoju, S. Roy, P. Parwekar, & L. Goel (Eds.), Proceedings of International Conference on Recent Trends in Computing (Vol. 341, pp. 19–49). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-7118-0_3

Saini, P., & Mishra, A. (2025). EFFECT OF ONLINE REVIEW ON SELECTION OF 5-STAR HOTEL. International Journal For Multidisciplinary Research, 7(2), 40376. https://doi.org/10.36948/ijfmr.2025.v07i02.40376

Saraswati, N. W. S., Ketut Gede Darma Putra, I., Sudarma, M., & Made Sukarsa, I. (2024). Enhance sentiment analysis in big data tourism using hybrid lexicon and active learning support vector machine. Bulletin of Electrical Engineering and Informatics,

13(5), 3663–3674. https://doi.org/10.11591/eei.v13i5.7807

Shehu, H. A., Sharif, Md. H., Sharif, Md. H. U., Datta, R., Tokat, S., Uyaver, S., Kusetogullari, H., & Ramadan, R. A. (2021). Deep Sentiment Analysis: A Case Study on Stemmed Turkish Twitter Data. IEEE Access, 9, 56836–56854. https://doi.org/10.1109/ACCESS.2021.3071393

Sreenivas, G., Murthy, K. M., Prit Gopali, K., Eedula, N., & H R, M. (2023). Sentiment Analysis of Hotel Reviews—A Comparative Study. 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), 1–9. https://doi.org/10.1109/I2CT57861.2023.10126445

Touahri, I. (2022). The construction of an accurate Arabic sentiment analysis system based on resources alteration and approaches comparison. Applied Computing and Informatics. https://doi.org/10.1108/ACI-12-2021-0338

Wang, C., Yang, X., & Ding, L. (2021). Sentiment classification based on weak tagging information and imbalanced data. Intelligent Data Analysis, 25(3), 555–570. https://doi.org/10.3233/IDA-205408

Zhu, J.-P., Niu, B., Cai, P., Ni, Z., Wan, J., Xu, K., Huang, J., Ma, S., Wang, B., Zhou, X., Bao, G., Zhang, D., Tang, L., & Liu, Q. (2024). *Towards Automated Cross-domain Exploratory Data Analysis through Large Language Models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2412.07214