

Stacking ensemble for multilingual aspect-based sentiment analysis of tourism reviews

Basworo Ardi Pramono^{1,2}, Rahmat Gernowo², Aghus Sofwan³

¹Department of Informatics Engineering, Faculty of Information and Communication Technology, Universitas Semarang, Semarang, Indonesia

²Doctor of Information Systems, Graduate School, Universitas Diponegoro, Semarang, Indonesia

³Department of Electrical Engineering, Faculty of Engineering, Universitas Diponegoro, Semarang, Indonesia

Article Info

Article history:

Received Jun 7, 2025

Revised Aug 4, 2026

Accepted Aug 12, 2026

Keywords:

Amenities, attractions, accessibility

Aspect-based sentiment analysis

Multilabel classification

Multilingual tourism reviews

Stacking ensemble

ABSTRACT

Multilingual tourism reviews often express opinions about multiple destination attributes within a single text, making aspect-level sentiment classification challenging. This study compares classical machine learning and ensemble methods for multilabel aspect-based sentiment analysis of amenities, attractions, and accessibility. The dataset comprises 13,085 Indonesian- and English-language Google Reviews labeled through a rule-based procedure. Reviews were represented using term frequency-inverse document frequency features and classified using logistic regression, multinomial naive Bayes, random forest, soft voting, adaptive boosting (AdaBoost), and stacking under a one-vs-rest framework. The stacking classifier combined logistic regression and random forest as base learners with logistic regression as the meta-learner. It achieved the best performance among the evaluated models, with a micro-F1 of 0.8626, macro-F1 of 0.7285, Hamming loss of 0.0633, and subset accuracy of 0.6825. These findings indicate that stacking effectively integrates complementary decision patterns and provides a reproducible framework for multilingual multilabel analysis of tourism reviews.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Basworo Ardi Pramono

Department of Informatics Engineering, Faculty of Information and Communication Technology

Universitas Semarang

St. Soekarno-Hatta, Semarang 50196, Central Java, Indonesia

Email: basworo@usm.ac.id

1. INTRODUCTION

Understanding travelers' sentiments through online reviews is increasingly vital for sustaining and enhancing the competitiveness of tourism destinations globally. User-generated content significantly shapes consumer decisions, impacting tourism businesses directly by influencing service reputation and quality [1], [2]. Thus, extracting precise insights from multilingual tourism reviews becomes crucial in the highly competitive international tourism market.

Aspect-based sentiment analysis (ABSA), focusing explicitly on amenities, attractions, and accessibility, enables a nuanced understanding of tourist experiences. Despite their potential, multilingual reviews pose distinct classification challenges due to language heterogeneity, grammatical variation, and cultural context, necessitating robust analysis techniques [3], [4]. However, the systematic evaluation of classical machine learning and advanced ensemble methods, especially stacking, for multilingual ABSA remains limited, creating significant knowledge gaps. Specifically, comprehensive comparative assessments across languages and cultures, as well as the less explored accessibility aspect, are scarce [5]–[7].

Prior research predominantly uses classical machine learning methods, such as logistic regression, naive Bayes, and random forests, for ABSA in tourism contexts, yielding reliable yet often suboptimal performance. These approaches have been reported for hotel review ABSA and other evaluative tourism content [8]. However, multilingual settings introduce additional linguistic variability and label ambiguity that remain insufficiently addressed in many existing studies [9], [10].

The literature further indicates an underrepresentation of ensemble methods, such as voting, adaptive boosting (AdaBoost), and stacking, in ABSA studies, despite their recognized advantages for improving predictive accuracy by combining the strengths of multiple classifiers [11]. Recent studies confirm that ensemble approaches can significantly improve classification robustness, especially crucial in multilingual and culturally diverse data [12], [13]. Moreover, existing research inadequately addresses accessibility within the amenities, attractions, and accessibility framework, thereby limiting comprehensive sentiment analysis in tourism reviews. Accessibility significantly impacts traveler satisfaction, particularly for those with disabilities, yet this crucial component is frequently neglected or superficially examined [6], [14].

This study hypothesizes that a stacking ensemble approach can significantly enhance multilingual ABSA accuracy compared to classical models and simpler ensemble methods. Thus, this research aims to comparatively evaluate these methods using a dataset of Indonesian and English reviews from Central Java, focusing specifically on the three aspects (amenities, attractions, and accessibility). By addressing identified knowledge gaps, the research contributes methodological clarity and improved predictive frameworks, ultimately aiding destination managers and developers in refining tourism recommendation systems.

2. METHOD

This study adopts a structured and thorough computational and experimental design to conduct a comparative evaluation and analyze the performance of a set of machine learning models. The models tested include classical category and ensemble approaches applied to the ABSA task in a multilingual context. This investigation focuses on analyzing digital tourism reviews, with an emphasis on three crucial aspects in the tourism domain: amenity, attraction, and accessibility. The set of methodologies implemented in this research includes several fundamental stages that are interconnected, starting from the acquisition and labeling of review data, followed by text preprocessing for normalization and cleaning, extraction of informative features from text, transformation of label format for compatibility with multilabel classification algorithms, segmentation of datasets into training and test datasets, implementation of multilabel classification architecture, model training process, evaluation of model performance using standard metrics, to the final stage of model selection with optimal performance. The overall methodological workflow is presented in Figure 1.

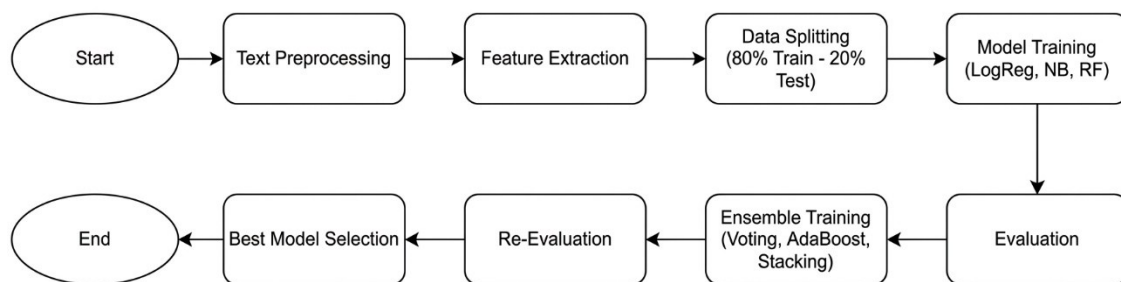


Figure 1. Overall workflow of the proposed research method

2.1. Data collection and filtering

The primary data source for this study is user reviews collected from the Google Reviews platform. This platform was chosen because it provides a large, diverse volume of review data covering a wide range of tourism destinations and is written in multiple languages by travelers with different backgrounds. Specifically, the dataset comprises 13,085 reviews from ten popular tourism destinations in Central Java, Indonesia. The selection of these destinations aims to obtain a representative variety of reviews from different types of tourist attractions.

Inclusion criteria were strictly applied during the data collection process to ensure the relevance and quality of the data to be analyzed:

- i) Review language: only reviews written in Indonesian and English language were included. This decision was based on the research's multilingual focus and the dominance of these two languages in the context of digital tourism in Indonesia.

- ii) Review length: each review must be at least 50 words long. These restrictions are applied to ensure that the review contains sufficient contextual information for in-depth analysis of aspects and sentiment and to avoid reviews that are too short and uninformative.
- iii) Relevance of aspects: the review must explicitly mention at least one of the three main aspects that the research focuses on: amenities (facilities and services), attractions (the main attraction of the destination), or accessibility (ease of reaching and moving around the destination).

After an initial collection and screening process based on the above criteria, reviews that meet the criteria are then analyzed at the clause level using a rule-based labeling approach, with the help of a structured keyword dictionary to identify appropriate aspects and sentiment polarities (positive, negative, neutral) [3]. The dataset used in this study is publicly available on Zenodo and contains labeled tourism reviews from Central Java, Indonesia, annotated using a rule-based approach [15]. The dataset includes multilingual review texts and the corresponding aspect sentiment labels used in the multilabel classification experiments.

2.2. Text preprocessing

Before feature extraction, the multilingual review text (stored in the review_clean column after the initial cleanup during data collection) undergoes several standard preprocessing steps to normalize further and clean it. The multilingual review text preprocessing process includes several important stages: normalization, tokenization, and noise removal. Normalization ensures consistency in text processing by converting entire characters to lowercase and handling spelling variations. Tokenization breaks down text into smaller units (tokens), which is especially important for languages like Indonesian and English because of their different syntactic structures. Noise removal eliminates irrelevant elements, such as punctuation and stop words, that can obscure the sentiment in user-generated content [16]. This stage is crucial for reducing noise and improving the quality of the features to be extracted for modeling.

2.3. Labeling scheme and label transformation

Labeling scheme and label transformation. The dataset was annotated using a rule-based scheme supported by a structured bilingual keyword dictionary for the three tourism aspects (amenities, attractions, and accessibility) and three sentiment polarities (positive, negative, and neutral). Each review was segmented into clauses to reduce sentiment mixing across aspects, and each clause was matched against the aspect dictionary to identify the relevant aspect category. Sentiment polarity was then assigned using polarity keywords and contextual rules (e.g., negation handling), producing clause-level aspect sentiment pairs.

To support multilabel learning at the review level, all clause-level pairs within the same review were aggregated into a set of composite labels in the format "aspect_polarity". Therefore, one review can have multiple labels simultaneously (e.g., amenity_positive and accessibility_negative). The final label set consists of 3 aspects × 3 polarities, for a total of 9 possible labels. The aggregated label sets were converted to a multi-hot representation using a MultiLabelBinarizer to construct the binary target matrix for one-vs-rest (OvR) training and evaluation.

2.4. Features extraction

For feature extraction, a term frequency-inverse document frequency (TF-IDF) vectorizer is used, a common technique in multilingual sentiment analysis [3]. TF-IDF highlights the importance of each word in the context of the entire corpus and adapts it to the frequency of common words, ensuring that more relevant features are given appropriate weight in the classification. TfidfVectorizer from the scikit-learn library was used to implement this research code. The max_features = 5000 parameter is applied, meaning only 5000 words with the highest TF-IDF score from the entire corpus will be considered features. This max_features also serves as a dimensionality reduction method, helping reduce model complexity and the risk of overfitting. The result of this feature extraction stage is a numerical feature matrix, denoted X. Each row of X corresponds to a review, and each column corresponds to a word feature with its TF-IDF weight.

2.5. Label transformation and data splitting

The sentiment-aspect labels resulting from the labeling process (as a list of tuples) are transformed into a multilabel binary format using the MultiLabelBinarizer from scikit-learn. Each unique combination of sentiment aspects (e.g., 'amenitas_positive') becomes a single binary target column. The datasets, which already included numerical features (X) and binary labels (y), were then split at an 80:20 ratio for training and test data. This ratio was chosen to balance sufficient learning and representative evaluation of unseen data [9].

2.6. Multilabel classification framework and tested models

This study formulates multilingual ABSA as a multilabel text classification problem, where a single review may contain multiple aspect sentiment pairs. Each review is represented by a set of composite labels formed by combining the aspect category and sentiment polarity (e.g., "aspect_polarity"), resulting in

3 aspects $\times$ 3 polarities = 9 possible labels. The label sets are binarized using MultiLabelBinarizer, producing a multi-hot target vector for each review.

For textual representation, TF-IDF vectorization is used on the cleaned review text (review_clean) with max_features = 5000 (default unigram setting). The dataset is split into training and test sets using an 80/20 ratio with a fixed random seed (random_state = 42). Because multilabel learning requires modeling each label as a binary decision, we employ a OvR strategy as a wrapper for all classifiers, enabling consistent comparison across baseline and ensemble approaches.

i) Baseline models:

- Logistic regression (OvR): LogisticRegression(max_iter = 1000, random_state = 42).
- Multinomial naive Bayes (OvR): MultinomialNB() as a probabilistic baseline suitable for TF-IDF features.
- Random forest (OvR): RandomForestClassifier(n_estimators = 100, random_state = 42).

ii) Ensemble models:

- Voting (OvR, soft Voting): VotingClassifier(voting = 'soft') with component learners logistic regression, multinomial naive Bayes, and random forest. Soft voting aggregates predicted probabilities to improve robustness.
- AdaBoost (OvR): AdaBoostClassifier (estimator = LogisticRegression (max_iter = 1000, random_state = 42), n_estimators = 50, random_state = 42) to iteratively emphasize misclassified samples.
- Stacking (OvR): StackingClassifier with base learners logistic regression and random forest, and logistic regression as the meta-learner (final_estimator = LogisticRegression()). We set passthrough = true so meta-learner can access both the original TF-IDF features and base-model predictions. Figure 2 illustrates end-to-end experimental setup used in this study, including preprocessing, label construction, feature extraction, model training, and multilabel evaluation.

The workflow starts from the Zenodo dataset and proceeds with text preprocessing to produce cleaned review texts. From the preprocessed data, two parallel processes are performed: i) a rule-based labeling procedure generates review-level multilabel targets, which are transformed into a multi-hot representation using MultiLabelBinarizer, and ii) TF-IDF vectorization (max_features = 5000) constructs the feature matrix. The data are then split into training and testing sets (80/20, random_state = 42) and trained under a OvR multilabel wrapper to enable a fair comparison between baseline models (logistic regression, multinomial naive Bayes, random forest) and ensemble approaches (voting, AdaBoost, and stacking). Finally, model performance is evaluated using micro- and macro-F1, precision, recall, Hamming loss, and subset accuracy, and the resulting metrics are compiled in the reported results tables.

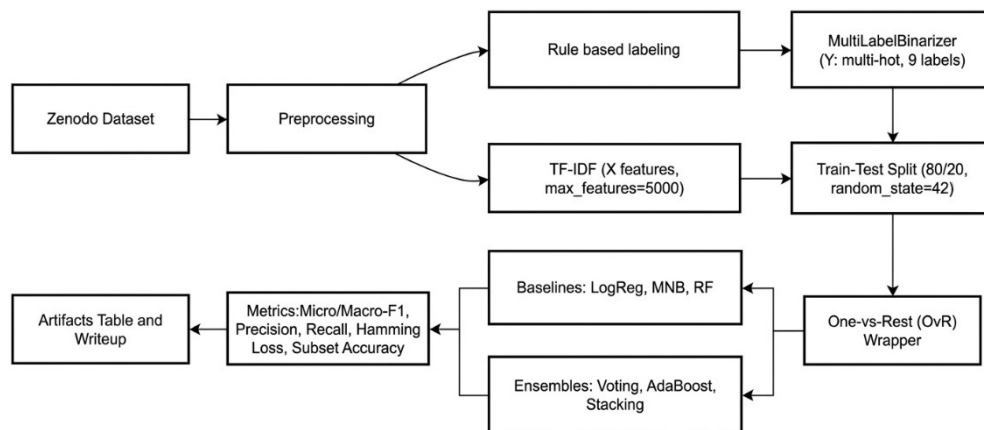


Figure 2. Experimental setup of the multilingual multilabel ABSA pipeline

2.7. Evaluation metrics

The evaluation used five metrics: F1-score (micro and macro), precision, recall, Hamming loss, and subset accuracy. This metric provides a comprehensive overview of the classification performance in multilabel tasks. The F1-score combines precision and recall, while Hamming loss measures prediction errors per label, and subset accuracy assesses overall label suitability [17]. Micro and macro averages yield different interpretations: micro averages describe overall performance, while macro averages equalize performance across classes, revealing potential imbalances [7].

In the context of multilabel classification for ABSA in multilingual tourism reviews, evaluation metrics such as precision, recall, F1-score (micro and macro), Hamming loss, and subset accuracy are essential for accurately assessing the model's performance. Precision measures the accuracy of the model's positive predictions [18]. On the other hand, recall measures the model's ability to capture all relevant instances [19]. F1-scores that combine precision and recall are particularly useful in multilabel tasks [20]. Hamming loss is the proportion of labels predicted to be incorrect [21]. Subset accuracy measures the proportion of label predictions that match the actual label. These metrics have been widely used in ABSA settings and have proven relevant for measuring the performance of sentiment analysis models and their resilience in handling multilabel data, especially in unbalanced classification [22].

Precision measures how accurately the model's positive predictions are. This explains the correct proportion of the prediction of the overall positive prediction, as shown in (1).

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

Recall measures the extent to which the model can capture all of the actual positive instances. This is important to avoid mistakes in skipping labels that should have been predicted, as shown in (2).

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

F1-score is a harmonic mean of precision and recall, providing a balanced assessment, especially when label distributions are imbalanced. There are two types: F1-micro and F1-macro. F1-micro calculates the total true positives, false positives, and false negatives in aggregate across all labels. The F1-macro calculates the F1 for each label individually and then averages it. This gives equal attention to each class, including the rare ones, as shown in (3).

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Hamming loss is a metric that measures the average number of misclassifications per label. In the context of multilabel classification, the Hamming loss is the proportion of incorrectly predicted labels (both false positives and false negatives) relative to the total number of labels. The lower the Hamming loss, the better the model's performance. A value of 0 means that all labels are correctly classified, as shown in (4).

$$Hamming\ loss = \frac{1}{n \times L} \sum_{i=1}^n \sum_{j=1}^L 1_{[y_{ij} \neq \hat{y}_{ij}]} \quad (4)$$

Where n is number of samples, L is number of labels per sample, and $y_{ij} \neq \hat{y}_{ij}$ is actual and predicted labels for the j -label in the i -I sample.

The accuracy subset is a metric that shows the percentage of a sample in which the predicted label is exactly the same as the actual label. This metric is very strict. If there is only one wrong label in a multilabel prediction, then all predictions are considered false for that sample. The higher the subset accuracy, the more precise the model's predictions are in handling label combinations, as shown in (5).

$$Subset\ accuracy = \frac{1}{n} \sum_{i=1}^n 1_{Y_i = \hat{Y}_i} \quad (5)$$

Where Y_i dan $\hat{Y}_i$ is actual and predicted label sets for the i th sample.

2.8. Ensemble training and re-evaluation

Once the baseline models are evaluated, the focus shifts to training and evaluating the assembly models: voting, AdaBoost, and stacking. The ensemble approach, particularly stacking, aims to combine predictions from basic models to improve generalization capabilities and the accuracy of final predictions. Evaluating the performance of ensemble stacking methods, such as voting, AdaBoost, and stacking, for multilingual ABSA in tourism reviews is critical given the complex nature of sentiment expressed across different languages and cultural contexts. Ensemble methods have been shown to improve classification accuracy by leveraging the collective strength of multiple models, thereby reducing the limitations associated with individual classifiers [23]. These methods are particularly beneficial in tourism, where sentiment about various aspects such as facilities, attractions, and accessibility can vary significantly across languages, making the challenges more pronounced [24], [25]. By integrating various algorithms, the ensemble stacking

Stacking ensemble for multilingual aspect-based sentiment analysis of tourism ... (Basworo Ardi Pramono)

method offers increased resistance to noise and better generalization to unseen data, which is especially important when dealing with the unbalanced datasets often encountered in sentiment analysis [26]. In addition, the ability of the stacked ensemble to combine predictions from different classifiers provides a more nuanced understanding of the sentiments expressed in multilingual reviews, leading to more reliable insights for stakeholders in the tourism industry [27], [28]. This, in turn, can drive progress in automated opinion mining and improve customer satisfaction through tailored recommendations based on accurate sentiment assessments [29]. The ensemble model was then re-evaluated using the same metric to compare its performance improvements relative to the baseline model and between fellow ensemble models [30].

2.9. Best model selection

The best models are selected based on the highest aggregate performance across all evaluation metrics. In this study, the stacking ensemble showed the best performance and was proposed as the most accurate and reliable model for ABSA on multilingual tour reviews [14], [29]. The secondary objective is to identify the best-performing models for tourism recommendation systems or automated opinion polling to support user and service provider decision-making. This research also helps bridge the gap in multilingual ABSA by emphasizing key aspects of tourism, such as amenities, attractions, and accessibility [1]. Ethical considerations are also taken into account, as the data is public, anonymous, and does not contain personal or sensitive information.

2.10. Technical implementation and visualization

The entire experimental stage, from data preprocessing to model evaluation, is implemented using Python (version 3.12.12). This study's main Python libraries include Pandas for data manipulation, scikit-learn for implementing machine learning models and evaluation metrics, and ast for early-stage label data transformation. All experiments were run using Python with the scikit-learn library for model training and evaluation, as well as for the application of classification metrics. Descriptive statistics are used to summarize the model's performance. For graphical visualization of model results in comparison, which supports more effective interpretation and delivery of data-based information, the Matplotlib and Seaborn libraries are used. Compute experiments run in a Google Colaboratory environment.

3. RESULTS AND DISCUSSION

3.1. Results of baseline and ensemble model evaluation

This study presents the results of evaluations of three baseline machine learning models, namely logistic regression, naive Bayes, and random forest, as well as three assembly models: voting classifier, AdaBoost classifier, and stacking classifier. The optimal configuration for the stacking classifier uses logistic regression and random forest as the base model (base learners) level 0, and logistic regression as the meta-learner (level 1), with the `passthrough = true` option. The entire model is implemented using the OvR classifier wrapper to handle the multilabel classification properties of this task.

Model performance was evaluated using a set of standard metrics for multilabel classification, including F1-score (micro and macro averaged), precision (micro averaged), recall (micro averaged), Hamming loss, and subset accuracy. These metrics provide a comprehensive view of the model's prediction accuracy, generalization capabilities, and error rate. Table 1 presents the comparative results of all the models tested on the test dataset. The values presented are the average performance obtained.

Table 1. ABSA model evaluation results based on multilabel metrics

Model	F1-micro	F1-macro	Precision	Recall	Hamming loss	Subset accuracy
Logistic regression	0.796	0.571	0.881	0.725	0.088	0.588
Naive Bayes	0.694	0.423	0.812	0.606	0.126	0.459
Random forest	0.831	0.590	0.878	0.789	0.076	0.640
Voting	0.800	0.524	0.885	0.729	0.086	0.601
AdaBoost	0.799	0.666	0.806	0.793	0.094	0.564
Stacking	0.863	0.729	0.885	0.841	0.063	0.682

3.2. Model comparison visualization

Figure 3 compares the F1-micro and F1-macro scores of all evaluated baseline and ensemble models. This graph visually compares the F1-score values (micro and macro) for all models tested. From this visualization, the superiority of the stacking ensemble in both F1-score metrics becomes visible compared to the other models.

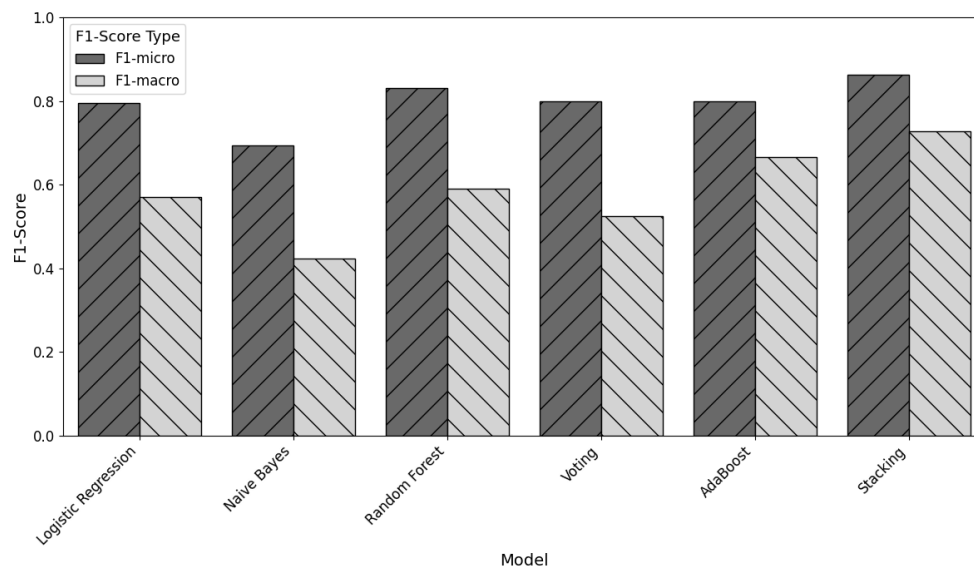


Figure 3. Comparison of F1-micro and F1-macro across the evaluated models

3.3. Description of results per model

A more detailed analysis of the results in Table 1 for each model is as follows.

3.3.1. Logistic regression

Logistic regression shows stable baseline performance for multilingual multilabel ABSA, providing a strong linear reference under TF-IDF features. The model achieves an F1-micro of 0.796 and an F1-macro of 0.571, with precision 0.881 and recall 0.725. While these results indicate reliable overall detection of common labels, the moderate macro-F1 suggests that performance is less uniform across all aspect sentiment labels, which is typical in multilabel datasets with label imbalance. The Hamming loss of 0.088 and subset accuracy of 0.588 further show that exact-match predictions remain challenging when multiple labels must be jointly predicted.

3.3.2. Naive Bayes

Multinomial naive Bayes performs worst among the evaluated models, reflecting its simplifying independence assumptions in complex multilingual review patterns. It records an F1-micro of 0.694 and an F1-macro of 0.423, with precision 0.812 and recall 0.606. The relatively low macro-F1 indicates that several labels, particularly less frequent aspect-sentiment combinations, are not consistently captured. This limitation is also reflected in the higher Hamming loss (0.126) and the lowest subset accuracy (0.459), showing difficulty in producing fully correct multilabel outputs.

3.3.3. Random forest

Random forest provides a strong non-linear baseline and performs competitively across the evaluated metrics. The model achieves an F1-micro of 0.831 and an F1-macro of 0.590, with precision 0.878 and recall 0.789. Compared to linear baselines, random forests benefit from modeling feature interactions, which can better capture heterogeneous wording in tourism reviews. Its Hamming loss decreases to 0.076 and subset accuracy increases to 0.640, indicating improved multilabel consistency and fewer per-label errors.

3.3.4. Voting classifier

The voting ensemble produces intermediate performance by combining complementary base learners through probability aggregation. It achieves an F1-micro of 0.800 and an F1-macro of 0.524, with precision 0.885 and recall 0.729. Although the ensemble stabilizes predictions relative to single learners, the macro-F1 remains lower than stronger ensembles, indicating that minority labels still receive limited gains. This pattern is consistent with its Hamming loss (0.086) and subset accuracy (0.601), which improve modestly but do not match the best-performing model.

3.3.5. AdaBoost classifier

AdaBoost improves performance by sequentially focusing on misclassified instances, which can be beneficial when some aspects of sentiment labels are harder to detect. The model achieves an F1-micro of 0.799 and a notably higher F1-macro of 0.666, with precision 0.806 and recall 0.793. The higher macro-F1

suggests that AdaBoost yields more balanced label-wise performance than several baselines, potentially improving weaker labels. However, its Hamming loss (0.094) and subset accuracy (0.564) indicate that exact multilabel matches remain difficult and that gains are not uniform across all evaluation criteria.

3.3.6. Stacking classifier

The stacking classifier consistently demonstrates the best overall performance across the major evaluation metrics. It achieves the highest F1-micro (0.8626) and F1-macro (0.7285), indicating both strong overall accuracy and balanced label-wise performance. In addition, it produces the lowest Hamming loss (0.0633) and the highest subset accuracy (0.6825), showing fewer label errors and better exact-match performance. These results suggest that meta-learning effectively combines complementary strengths of the base learners (logistic regression and random forest) through a logistic regression meta-learner.

3.4. Discussion

The results showed that the stacking ensemble model consistently outperformed classical machine learning models and other ensemble methods in multilabel classification of multilingual tour reviews. This is reflected in the F1-score micro value of 0.8626 and the higher F1-score macro of 0.7285, as well as the lowest Hamming loss of 0.0633 and the highest subset of accuracy reaching 0.6825. In comparison, the strongest baseline model, random forest, achieved a micro F1-score of 0.8313 and a macro-F1-score of 0.5898, with a Hamming loss of 0.0756 and a subset accuracy of 0.6404. Other ensemble models, such as AdaBoost, achieve micro F1-scores of 0.7991 and macro F1-scores of 0.6655, while voting achieves micro F1-scores of 0.7999 and macro F1-scores of 0.5241. These findings reinforce the premise that the integration of multiple basic models (in this case, logistic regression and random forest as base learners, with logistic regression as the meta-learner) through a stacking approach can improve the generalization and accuracy of predictions, especially in the context of complex and diverse data such as travel reviews.

In this study, the results showed that the stacking ensemble method significantly outperformed classical machine learning models such as logistic regression (F1-micro: 0.7958), naive Bayes (F1-micro: 0.6942), and random forest (F1-micro: 0.8313), as well as other ensemble methods such as voting (F1-micro: 0.7999; Hamming loss: 0.0861; subset accuracy: 0.6007) and AdaBoost (F1-micro: 0.7991; Hamming loss: 0.0941; subset accuracy: 0.5640) in various key evaluation metrics. Specifically, the stacking ensemble not only achieves the highest F1-score (micro: 0.8626, macro: 0.7285) but also the lowest Hamming loss (0.0633) and the highest subset accuracy (0.6825). This superior performance of the stacking ensemble demonstrates its ability to combine predictions from multiple classifiers, which is particularly beneficial in the multilingual landscape of tourism reviews, where sentiment expressions can vary widely [16]. These findings are in line with previous literature [11], [12], which highlight the effectiveness of ensemble techniques in sentiment classification tasks, and support the findings of [3], [8] regarding the benefits of stacking for analyzing data across dialects and languages.

By extending these results into the context of multilingual ABSA, the implications of using stacking ensemble become significant. Its ability to achieve higher accuracy of aspect and sentiment predictions (evidenced by a micro F1-score of 0.8626) can improve the accuracy and reliability of sentiment analysis on tourism-related user-generated content. This, in turn, supports better decision-making for industry stakeholders [31], [32]. This, in turn, supports better decision-making for industry stakeholders. For example, it can help identify a destination's specific strengths and weaknesses by analyzing the aspects most frequently discussed, along with their sentiment polarities.

However, it should be noted that while the stacking model performs well with the TF-IDF feature, rule-based labelling approaches used in the early stages of data collection or as a comparator still face challenges, including potential inaccuracies in sentiment classification due to contextual language nuances. Future research may explore integrating transformer-based models such as bidirectional encoder representations from transformers (BERT) or cross-lingual language model based RoBERTa (XLM-R) with this proven stacking architecture. Transformer models, which have demonstrated higher accuracy through contextual understanding in sentiment analysis tasks [29], [33], potentially further improve performance, especially when combined with the generalization capabilities of stacking. In addition, the importance of multilabel evaluation metrics such as F1-score per label (which can also be calculated from your model, e.g., for F1_accessibility_positive, F1_amenities_negative), Hamming loss, and subset accuracy is becoming increasingly clear. The metric is important for comprehensively analyzing user opinions in a multilingual ABSA system, not only supporting an in-depth understanding of sentiment related to various aspects of tourism but also providing important insights into user preferences and experience [30], [34].

Thus, this study not only fills a methodological gap in the multilingual ABSA literature by demonstrating the superiority of the stacking ensemble on the multilingual tourism review dataset in Indonesia, but also provides an important foothold for further exploration. The success of stacking ensemble

with specific configurations (logistic regression and random forest as base learners, and logistic regression as meta-learners) in achieving subset accuracy of 0.6825 (meaning that 68.25% of reviews are perfectly classified for all sentiment labels) and Hamming loss 0.0633 (average error per label is very low) became benchmarks for the development of an opinion analysis and destination recommendation system based on user review data more advanced and accurate.

4. CONCLUSION

This study benchmarked baseline and ensemble machine learning methods for multilingual multilabel ABSA on tourism reviews covering amenities, attractions, and accessibility. Using TF-IDF features and an OvR multilabel framework, the results show that stacking provides the strongest and most consistent performance across F1-micro, F1-macro, Hamming loss, and subset accuracy, outperforming logistic regression, naive Bayes, random forest, voting, and AdaBoost. The findings indicate that meta-learning effectively combines complementary model strengths and yields better label-wise balance, which is crucial for extracting reliable aspect-sentiment signals from bilingual tourism data. In practice, these signals can help tourism stakeholders prioritize service improvements aligned with user feedback across the three aspect categories. Future work should validate the approach under additional robustness settings (e.g., cross-validation and external test datasets) and compare it against multilingual transformer-based encoders to assess generalization beyond TF-IDF representations.

ACKNOWLEDGMENTS

The authors thank Universitas Semarang and Universitas Diponegoro for the institutional, computational, and academic support provided during this research.

FUNDING INFORMATION

This research received institutional funding from Universitas Semarang. The funding was provided without a specific grant or contract number.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Basworo Ardi Pramono	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓
Rahmat Gernowo	✓	✓		✓	✓		✓			✓		✓		
Aghus Sofwan	✓	✓		✓			✓			✓		✓		

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**xperimentation

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

INFORMED CONSENT

This section not applicable. This study analyses publicly available online reviews. No human subjects were recruited, and no personally identifiable information was collected, stored, or reported.

ETHICAL APPROVAL

This section not applicable. The study is based on publicly available online reviews. All user identifiers (e.g., usernames and profile links) were removed during preprocessing, and only anonymized text data were analyzed. Therefore, institutional ethical approval was not required.

DATA AVAILABILITY

The dataset analyzed in this study is publicly available on Zenodo at <https://doi.org/10.5281/zenodo.15517727>.

REFERENCES

- [1] Z. A. -Moud, H. V. -Nejad, and J. Sadri, "Tourism recommendation system based on semantic clustering and sentiment analysis," *Expert Systems with Applications*, vol. 167, Apr. 2021, doi: 10.1016/j.eswa.2020.114324.
- [2] S. Ghosal and A. Jain, "Weighted aspect based sentiment analysis using extended OWA operators and Word2Vec for tourism," *Multimedia Tools and Applications*, vol. 82, no. 12, pp. 18353–18380, May 2023, doi: 10.1007/s11042-022-13800-4.
- [3] D. Ariyus, D. Manongga, and I. Sembiring, "Enhancing sentiment analysis of Indonesian tourism video content commentary on TikTok: a fasttext and Bi-LSTM approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18020–18028, Dec. 2024, doi: 10.48084/etasr.8859.
- [4] A. M. -Ortiz, S. S. -Bernal, and A. O. -Barca, "Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector," *Information Technology & Tourism*, vol. 21, no. 4, pp. 535–557, Dec. 2019, doi: 10.1007/s40558-019-00155-0.
- [5] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Computer Science Review*, vol. 41, Aug. 2021, doi: 10.1016/j.cosrev.2021.100413.
- [6] C.-H. Ku, Y.-C. Chang, and Y. Wang, "How to strategically respond to online hotel reviews: a strategy-aware deep learning approach," *Information & Management*, vol. 61, no. 5, Jul. 2024, doi: 10.1016/j.im.2024.103970.
- [7] N. Habbat and H. Nouri, "Unlocking travel narratives: a fusion of stacking ensemble deep learning and neural topic modeling for enhanced tourism comment analysis," *Social Network Analysis and Mining*, vol. 14, no. 1, Apr. 2024, doi: 10.1007/s13278-024-01256-3.
- [8] A. Onan, "Mining opinions from instructor evaluation reviews: a deep learning approach," *Computer Applications in Engineering Education*, vol. 28, no. 1, pp. 117–138, Jan. 2020, doi: 10.1002/cae.22179.
- [9] A. Ghorbanali, M. K. Sohrabi, and F. Yaghmaee, "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks," *Information Processing & Management*, vol. 59, no. 3, May 2022, doi: 10.1016/j.ipm.2022.102929.
- [10] C. Yu, X. Zhu, B. Feng, L. Cai, and L. An, "Sentiment analysis of Japanese tourism online reviews," *Journal of Data and Information Science*, vol. 4, no. 1, pp. 89–113, Feb. 2019, doi: 10.2478/jdis-2019-0005.
- [11] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar, "A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13407–13461, Nov. 2023, doi: 10.1007/s10462-023-10472-w.
- [12] A. A. Al Shamsi and S. Abdallah, "Ensemble stacking model for sentiment analysis of Emirati and Arabic dialects," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, Sep. 2023, doi: 10.1016/j.jksuci.2023.101691.
- [13] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.014.
- [14] M. Paolanti *et al.*, "Tourism destination management using sentiment analysis and geo-location information: a deep learning approach," *Information Technology & Tourism*, vol. 23, no. 2, pp. 241–264, Jun. 2021, doi: 10.1007/s40558-021-00196-4.
- [15] B. A. Pramono, "Multilingual aspect-based sentiment analysis dataset of tourism reviews in Indonesia (labeled with Gemini 2 LLM)," *Zenodo*, 2025, doi: 10.5281/zenodo.15517727.
- [16] A. Tannirisevdi, G. Öztürk, and A. C. Öztürk, "A supervised data mining approach for predicting comment card ratings," *International Journal of Contemporary Hospitality Management*, vol. 34, no. 5, pp. 1823–1853, Apr. 2022, doi: 10.1108/IJCHM-05-2021-0675.
- [17] S. Huang, L. J. Liang, and H. C. Choi, "How we failed in context: a text-mining approach to understanding hotel service failures," *Sustainability*, vol. 14, no. 5, Feb. 2022, doi: 10.3390/su14052675.
- [18] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.01406116.
- [19] B. Harjo, Muljono, and R. Abdullah, "Attention-based sentence extraction for aspect-based sentiment analysis with implicit aspect cases in hotel review using machine learning algorithm, semantic similarity, and BERT," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 3, pp. 189–200, Jun. 2023, doi: 10.22266/ijies2023.0630.15.
- [20] C.-C. Wu, C.-W. Tsai, F.-E. Wu, C.-H. Chiang, and J.-C. Chiou, "Plantar pressure-based gait recognition with and without carried object by convolutional neural network-autoencoder architecture," *Biomimetics*, vol. 10, no. 2, Jan. 2025, doi: 10.3390/biomimetics10020079.
- [21] M. Zhao *et al.*, "Establishment of a Daqu grade classification model based on computer vision and machine learning," *Foods*, vol. 14, no. 4, Feb. 2025, doi: 10.3390/foods14040668.
- [22] L. Yao *et al.*, "Rib fracture detection system based on deep learning," *Scientific Reports*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-03002-7.
- [23] S. Anis, S. Saad, and M. Aref, "Sentiment analysis of hotel reviews using machine learning techniques," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020 (AISI 2020)*, Cham, Switzerland: Springer, 2021, pp. 227–234, doi: 10.1007/978-3-030-58669-0_21.
- [24] M. B. Anley, G. S. Negashe, and A. Y. Ayalew, "Opinion mining of tourists' sentiments: towards a comprehensive service improvement of tourism industry," in *2023 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Jun. 2023, pp. 1–6, doi: 10.1109/CIVEMSA57781.2023.10231013.
- [25] S. Sagnika, A. Pattanaik, B. S. P. Mishra, and S. K. Meher, "A review on multilingual sentiment analysis by machine learning methods," *Journal of Engineering Science and Technology Review*, vol. 13, no. 2, pp. 154–166, Apr. 2020, doi: 10.25103/jestr.132.19.
- [26] M. K. Anam, M. B. Firdaus, F. Suandi, Lathifah, T. Nasution, and S. Fadly, "Performance improvement of machine learning algorithm using ensemble method on text mining," in *2024 International Conference on Future Technologies for Smart Society (ICFTSS)*, Aug. 2024, pp. 90–95, doi: 10.1109/ICFTSS61109.2024.10691363.

- [27] M. Á. Á. -Carmona, R. Aranda, R. G. -Rodríguez, A. Y. R. -González, and A. P. L. -Monroy, "A combination of sentiment analysis systems for the study of online travel reviews: many heads are better than one," *Computación y Sistemas*, vol. 26, no. 2, Jun. 2022, doi: 10.13053/cys-26-2-4055.
- [28] J. Kazmaier and J. H. V. Vuuren, "The power of ensemble learning in sentiment analysis," *Expert Systems with Applications*, vol. 187, Jan. 2022, doi: 10.1016/j.eswa.2021.115819.
- [29] Y. Liu and J. Li, "An online hotel selection method with three-dimensional analysis of reviews' helpfulness," *International Journal of Fuzzy System Applications*, vol. 13, no. 1, pp. 1–25, May 2024, doi: 10.4018/IJFSA.343490.
- [30] S. Smetanin, "The applications of sentiment analysis for russian language texts: current challenges and future perspectives," *IEEE Access*, vol. 8, pp. 110693–110719, 2020, doi: 10.1109/ACCESS.2020.3002215.
- [31] S. Anis, S. Saad, and M. Aref, "A survey on sentiment analysis in tourism," *International Journal of Intelligent Computing and Information Sciences*, vol. 20, no. 1, pp. 1–15, Jun. 2020, doi: 10.21608/ijicis.2020.106309.
- [32] D. Zhu, R. Jing, Q. Guo, D. Zhang, and F. Wan, "Sentiment analysis of tourism review text combined with bert-bilstm and attention mechanism," *Journal of Computational Methods in Sciences and Engineering*, vol. 24, no. 3, pp. 1605–1615, Jun. 2024, doi: 10.3233/JCM-247135.
- [33] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," *Information Processing & Management*, vol. 58, no. 5, Sep. 2021, doi: 10.1016/j.ipm.2021.102656.
- [34] D. Guidotti, L. Pandolfo, and L. Pulina, "Discovering sentiment insights: streamlining tourism review analysis with large language models," *Information Technology & Tourism*, vol. 27, no. 1, pp. 227–261, Mar. 2025, doi: 10.1007/s40558-024-00309-9.

BIOGRAPHIES OF AUTHORS



Basworo Ardi Pramono    is a doctoral candidate in the Doctoral Program of Information Systems at Universitas Diponegoro, Semarang, Indonesia. He received his S.T. (Electrical Engineering) from Universitas Negeri Semarang in 2006 and his M.T. (Electrical Engineering) from Institut Teknologi Bandung in 2011. He is currently a lecturer at the Faculty of Technology and Informatics Communication, Universitas Semarang. His research interests include artificial intelligence, natural language processing, decision support systems, and their applications in tourism and hospitality information systems. He has actively published articles in national and international journals and presented his work at academic conferences. His recent research focuses on integrating machine learning and NLP for aspect-based sentiment analysis and intelligent recommender systems in the tourism sector. He can be contacted at email: basworo@usm.ac.id.



Rahmat Gernowo    is currently a professor in the field of atmospheric physics models at Universitas Diponegoro in Semarang, Central Java. His current research focuses on the information system for disasters in the tropics. Several publications related to models for disaster information systems have been published in journals indexed by Scopus and Web of Science. He has also conducted international collaborative research with several national universities, funded by the UK-affiliated degrees modeling fund, in the solar radiation modeling for climate change studies field. Registered as a member of the association of organizations: Indonesian Computer Higher Education Association (APTIKOM), Indonesian Association of Geophysicists (HAGI), Indonesian Physicists Association (HFI), and Indonesian Society of Hydrology (MHI). He can be contacted at email: rahmatgernowo@lecturer.undip.ac.id.



Aghus Sofwan    is currently an associate professor in the field of Information Technology at Universitas Diponegoro in Semarang, Central Java, Indonesia. He earned his Ph.D. in Computer Engineering from King Saud University, where he specialized in data communication using information technology. His current research focuses on the implementation of artificial intelligence in information technology. Several publications related to the implementation of artificial intelligence have been published in journals or presented in international conferences indexed by Scopus and WoS. He is registered as a member of the association of organizations, such as IEEE, and an executive professional engineer in The Institution of Engineers Indonesia (PII). He can be contacted at email: asofwan@elektro.undip.ac.id.