Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Rethinking the role of three-star ratings: Handling inconsistency in indonesian tourism reviews using cost-sensitive learning XGBoost

Muhammad Nur Hikmah Ramadhan[1], Akhmad Syaifuddin[2], Ristu Saptono[3]
[1, 2, 3]Department of Informatics, Universitas Sebelas Maret, Indonesia

Article Info

Article history:

Received June 5, 2026
Revised June 19, 2026
Accepted June 24, 2026

Keywords:

Class restructuring
Cost-sensitive learning
eXtreme gradient boosting
FastText
Rating prediction

ABSTRACT

Indonesian people often give contrary online reviews, for instance ratings that do not always match the actual feelings. This can make it difficult for tourism sector, such as Solo Safari, to handle complaints and improve service quality. Consequently, this study has formulated a rating prediction model using the XGBoost algorithm with a dataset of 2,047 reviews that have been relabeled. The best model, the FastText-XGBoost with Cost-Sensitive Learning, signify that it performs quite well with an average difference the prediction is only 0.04 points from the actual rating. However, the results are not optimal because the meaning of the review text still feels fuzzy even though being relabeled. This problem arises because reviewers tend to position sentiment very positively or negatively in the moderate category, so the perimeters between classes become less clear. This study then proposed an extreme class restructuring by simplifying the category by removing the 3-star rating. This method can increase the accuracy of the model to 0.9440 and clarify the category limits. Therefore, the model can be used on the service dashboard to help Solo Safari management respond to critical feedback faster.

Corresponding Author:

Muhammad Nur Hikmah Ramadhan,
Department of Informatics,
Universitas Sebelas Maret,
Kentingan, Jl. Ir. Sutami No. 36A, Jebres, Jebres District, Surakarta City, Central Java 57126
Email: nurhikmah.2411@student.uns.ac.id
https://doi.org/10.52465/joscex.v7i3.159

1. INTRODUCTION

The percentage of consumers who share online reviews is increasing enormously, making them increasingly reliant on them [1]. In the tourism sector, the high reliance on online reviews stems from consumers (in this case, tourists) who use reviews written by other travellers to help them plan their travel. This behaviour is driven by the convenience of online reviews, which allow travellers to evaluate multiple tourist attractions at once on a single platform [2]. Especially now that platforms like Google Maps offer a variety of features that make it easier for them to express themselves or share online reviews [3]. The ease of access and high volume of interaction on these platforms make them key tools for gauging public reaction, especially when a destination undergoes a large-scale transformation. This phenomenon occurred at one of the tourist attractions in Surakarta, Indonesia, namely Solo Safari. Following its revitalisation and official opening on January 27, 2023 [4], this tourist attraction has seen a massive increase in visitor numbers [5]. This massive

influx of tourists will naturally be directly proportional to the boost in online reviews on the Solo Safari Google Maps page. The massive growth in online reviews on Solo Safari has made it an essential data source for evaluating facility quality and public happiness with the new management.

Although online review data is, in actuality, quite abundant, efforts to explore it are often hampered. The challenge is the discrepancy between the actual sentiment in the review text and the ratings given by travelers [6], [7]. This phenomenon is more generally known as Text-Rating Review Discrepancy (TRRD) [8]. A rating prediction procedure can address this annoyance by predicting ratings that reflect the actual emotional content of travelers' reviews [6]. In the initial stages of the procedure, the review must be validated through an annotation process conducted by several annotators to ensure the dataset is suitable and to identify inconsistent reviews early [8]. After several annotators finalized the annotation process, the agreement among the annotations was first assessed using Fleiss' Kappa [9]. The annotation process then resumed by choosing the final labels for the annotated results using a majority-vote scheme, after which any inconsistent entries in the final dataset were eliminated. A dataset that has been validated through prior objective testing will first undergo a preprocessing phase to regularize the textual data. This phase includes normalizing Indonesian slang, as well as diminishing the use of words that are difficult for the rating prediction model to determinate (out-of-vocabulary) [10]. The purpose of this step is to ensure that the review text is ready for vectorization and can be optimally comprehended by machine learning (ML) algorithms.

One of the many commonly used text representation methods is Term Frequency-Inverse Document Frequency (TF-IDF). Several studies have employed this method, combining it with classification or prediction algorithms such as XGBoost [11], [12]. The combination of the two in these studies produced fairly good results. Unfortunately, TF-IDF has been found to fail to capture the deep semantic meaning of text, so several subsequent studies have suggested using word embeddings such as Word2Vec and FastText [13]. In that study, word embeddings such as word2vec outperformed TF-IDF, and FastText achieved the best results of the two. Based on the results of this study, in theory, FastText is indeed superior to the other three methods because it can represent text at the subword level (character n-grams). As a result, this method can handle abbreviations, typos, and out-of-vocabulary (OOV) words [14]. There is even a study that validates its performance when combined with XGBoost, though it used FastText with TF-IDF to achieve optimal results [15].

Once the text representation was obtained, a well-known rating prediction model, such as XGBoost, was applied to the dataset [12]. However, public review datasets often face the problem of an unbalanced distribution across classes (imbalanced datasets), which can reduce algorithm performance. Although oversampling techniques such as ADASYN are commonly used to synthesize data for minority classes [16], the Cost-Sensitive Learning approach is recognized as an alternative to oversampling methods because it prevents the manipulation or loss of original information in the dataset [12], [17]. This technique works by directly imposing a larger penalty on the algorithm for each prediction error in the minority class.

Some challenges in this study, such as computational processes, can be overcome via text representation and effective data balancing. However, a study spotlights a simple fact, even completed annotation procedures remain liable to the annotator's subjectivity [18]. The high level of ambiguity in the moderate or neutral category is the essential cause of this problem. A study also provides a comprehensive finding indicating that more than 40% of online reviews contain contradictory emotions. Most of the sentiment polarity issues were found in the neutral class [19]. This phenomenon occurs because users' emotional expressions are frequently stronger or more powerless than neutral. As a effect, this moderate class acts as a source of noise because it lacks clear sentiment edges and can directly pollute model performance [20]. Therefore, to address this noise, removing 3-star reviews can be proposed as an analytical strategy. In fact, according to several studies, these have been shown to improve the accuracy of algorithmic predictions [20], [21]. Therefore, this study will implement this strategy, compare feature extraction techniques, and compare data balancing techniques to produce a representative and optimal model for evaluating tourist attractions.

2. METHOD

The methodology in this study was carefully prepared into four complementary phases to deliver an objective evaluation process free of data leakage. A graphic representation of the research workflow is illustrated in Figure 1, where the four steps are: data collection and initial data filtering; annotation and consensus; text preprocessing; and, finally, analysing and comparing all modelling scenarios and conducting class restructuring experiments.

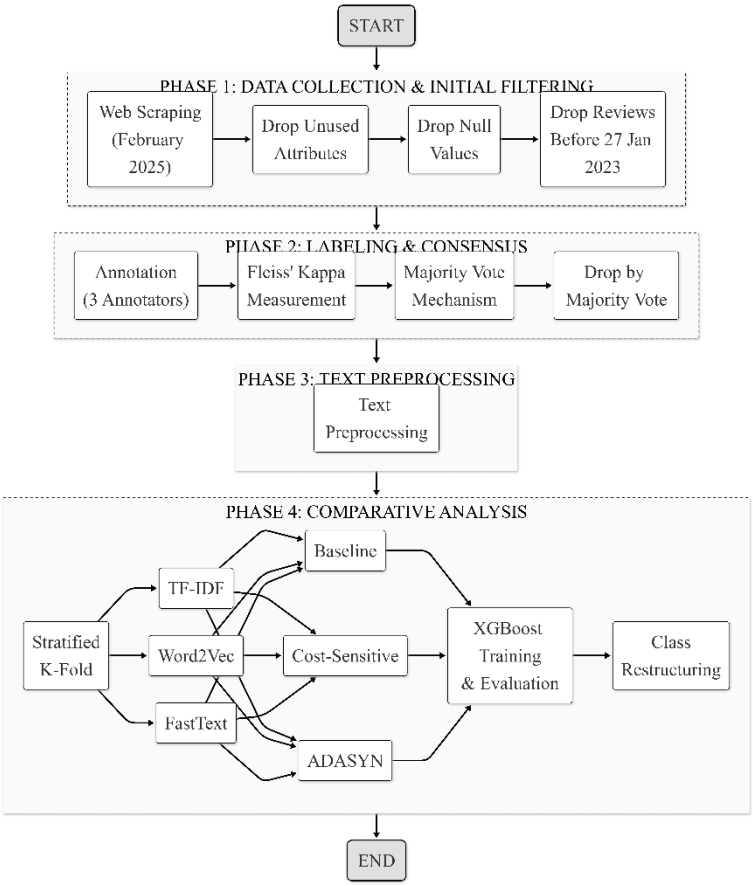

Figure 1. Research flow

Data Collection and Labelling

To begin this study, on February 27, 2025, the researchers collected data using the open-source Google Maps Scraper Application Programming Interface (API) on the Apify platform [22]. From that platform, the researchers obtained 10,000 Google Maps online traveler reviews for the Solo Safari destination. Unfortunately, this dataset still contains some reviews that are not actually intended for Solo Safari, but for the previous name, which is more specific, reviews posted before January 27, 2023. For the prevention of this study dataset from covering the management period of Solo Safari before its revitalization, especially reviews published before January 27, 2023, this study does not include such reviews. The researchers also selected several data rows that lacked review text for exclusion, resulting in a total of 4,342 records to be labeled to identify consistent and inconsistent reviews by three independent evaluators with academic backgrounds in computer science. The agreement between the annotated results is calculated using Kappa Fleiss and yields a value of 0.8274 indicating a "Almost Perfect Agreement" [9], [23]. These results are indirectly indicative of this data worthy of further analysis. To objectively determine the dataset, the researchers used the annotators' results, which were combined and evaluated using a majority voting scheme. In this voting process, the rule is that if two annotators have labeled the data with comments indicating inconsistent reviews, the researchers will remove it. If two annotators label the data as valid, it will be retained. Overall, this majority-vote process will eliminate much of the invalid and noisy data. The final result of this process was that only 2,214 valid reviews from Solo Safari visitors remained, ready to proceed to the next stage.

Text Preprocessing

The validated data is then split, and textual information is not extracted first. This is because the textual data will first undergo a series of preprocessing steps, including case folding, removal of duplicate or repeated characters, removal of emojis, removal of punctuation marks, normalization, and removal of stopwords [10], [12], [24]. To standardize or address slang, a specialized dictionary was used in this study [10], [25]. It should be noted that when removing stopwords, the researchers also created a separate dictionary of negation terms to preserve all negation words during the stopword removal process. This adjustment was justified because removing those negative words could alter the sentiment polarity of the reviews and affect their emotional context [12], [26], retaining them also helped the model map the sentiment orientation of the reviews more accurately. In this study, the author performed a process known as stemming, because affixes in Indonesian vocabulary often carry specific sentiment connotations that are essential to predictive analysis [27].

Finally, the researchers removed reviews consisting of fewer than three words, as retaining them would not provide sufficient information. The result of this series of processes is 2,047 cleaned review data points ready to be converted into vectors.

Feature Extraction and Imbalance Handling

After a lengthy preprocessing phase, the dataset was divided into 5 parts, maintaining the proportions of each class as in the original dataset, using Stratified k-Fold Cross-Validation. This splitting process occurs before the text reviews are converted into numerical vectors, as it also divides the dataset into training and test sets. This splitting is also intended to ensure a more stable evaluation and prevent data leakage [28], [29]. Once the dataset is divided into its constituent parts, text reviews are extracted and vectorized using feature extraction techniques. All model parameters and feature extraction methods were determined through iterative experimental testing to align with the characteristics of the review data. Since this stage also involves comparative experiments, three feature extraction methods were applied: the statistical TF-IDF approach (N-grams 1–2, minimum count 2) [11], [12], and word embedding methods such as Word2Vec and FastText [13], [15], [30]. The results of the final experiment on the Word2Vec parameter indicated that the best performance on this dataset was obtained with the Skip-Gram architecture, 200 vector dimensions, 5 context window sizes, minimum count 2, and 50 training epochs. This setting is also used in FastText. However, in FastText, the model is reconfigured to recognize features down to the subword level with an emphasis on N-Gram of 2. Once the vector results are obtained, the XGBoost algorithm process begins. Prior to that, this study also compared two approaches separately, namely ADASYN oversampling [16] and Cost-Sensitive Learning to balance class distributions [12], [17].

Model Training and Evaluation

The data that has been extracted using TF-IDF, Word2Vec, and FastText is further processed with the XGBoost algorithm [12], [31]. This algorithm uses a maximum tree depth of 5, a learning rate of 0.05, a subsample of 0.8, and regularization parameters of alpha 0.1 and lambda 2.0. XGBoost can be used to compare feature extraction techniques and data balancing methods to get the best combination. Each model was evaluated in each iteration, then the three highest-performing models were tested using the nonparametric Wilcoxon Signed-Rank Test with a $p <$ value of 0.05 [32], [33]. The purpose of this test is to ensure that the difference in performance between models is significant and not accidental. The final result will determine the most optimal model. However, in this case, a class restructuring will be implemented and further evaluated to determine whether conflicting sentiments actually exist. If necessary, this study then performs class reduction to eliminate the noisiest classes and analyze them further [19], [20], [21]. It then compares the results with those from the best 5-class experiment. The results of all tests will also be evaluated from a business perspective.

3. RESULTS AND DISCUSSIONS

Before the dataset was ready for training with XGBoost, in accordance with the research protocol, the researchers preprocessed it by clearing rows with empty columns and excluding all reviews posted before January 27, 2023. The annotators then labeled the dataset as “valid” or “drop.” At this stage, all data that had been removed was approved by a majority vote, resulting in a total of 2,214 reviews being retained. Of those 2,214 reviews, 2,047 were retained after preprocessing. An immediate analysis of the dataset revealed that the data distribution was highly skewed. The majority category is dominated by 5-star reviews (1,151), in stark contrast to the minority category, the neutral 3-star category, with only 102 reviews. An overall breakdown of the data distribution after preprocessing is shown in Table 1. After this process, the researchers applied 5-fold stratified cross-validation to ensure the model could adapt to differences in data distribution.

Table 1. Class distribution in the dataset after preprocessing

Star Rating	Number of Reviews	Percentage
1	318	15.54%
2	148	7.23%
3	102	4.98%
4	328	16.02%
5	1151	56.23%

Impact of Annotator Validation and Negation Word Handling

The first-stage evaluation was conducted by comparing the model's training results on a dataset of 4,342 initial reviews with the results on the annotated dataset. The results showed a significant difference. Models trained on a noisy dataset with inconsistent reviews achieved only 52–55 percent accuracy. These results stand in stark contrast to those of models trained on annotated datasets, which showed an improvement of 73–75 percent. This condition proves that Solo Safari reviews are full of noise and urgently need relabeling. Specifically, the complete results of the comparison of the three feature extraction methods with the baseline XGBoost model, applied to the dataset both before and after annotation, are presented in Table 2.

Table 2. Comparison of models before and after annotation

Method	Accuracy (Before Annotation)	F1-Score (Before Annotation)	Accuracy (After Annotation)	F1-Score (After Annotation)
XGBoost + TF-IDF	0.5297	0.5060	0.7365	0.7113
XGBoost + Word2Vec	0.5502	0.5298	0.7483	0.7304
XGBoost + FastText	0.5415	0.5229	0.7542	0.7410

The next experiment was designed to analyze the impact of modifying the stop-word dictionary, specifically by retaining negation words to preserve the integrity of the text representation. As is common in many preprocessing steps, negation words such as "Tidak" or "Kurang" are often automatically removed because they are considered to have no particular weight; however, removing them can distort the review's sentiment. The researchers then conducted a direct analysis of the Solo Safari dataset (ID 2190). They found that the review contained several negative statements, such as "panas, ga ada angin sama sekali, pertunjukan umum kurang bagus karna satpam kurang sopan...". In this review, if the word "ga, kurang" is omitted then the sentiment directly changes to a different meaning. The review will read: "panas, ada angin, pertunjukan umum bagus karna satpam sopan". Sentiments that initially have a negative tone can turn into positive. This change in sentiment also reverses direction, where mathematically the direction of the vector changes causing the model to misclassify. So that this explanation is not indicated based only on one data, the researcher shows a comparison of the use of negation shown in Table 3.

Table 3. Impact of negation word handling on xgboost performance

Method	Accuracy (Without Negation Handling)	Accuracy (With Negation Handling)	Delta (Δ)
XGBoost + TF-IDF	0.7365	0.7523	+0.0158
XGBoost + Word2Vec	0.7483	0.7660	+0.0177
XGBoost + FastText	0.7542	0.7684	+0.0142

Comparison of Feature Extraction and Distribution Inequality

This study tested TF-IDF, Word2Vec, and FastText separately to determine the most effective feature extraction method. The test results in Table 3 indicate that FastText performs best with an average accuracy of 0.7684 and an F1 score of 0.7571. This performance is influenced by FastText's ability to understand word representation up to the subword level so that it is able to produce vectors better than other methods. This capability makes it more adaptable at reconstructing words that contain typos, abbreviations, slang, or even terms outside its vocabulary. The overall results of this test are shown in Table 4.

Table 4. Comparison of feature extraction method performance

Method	Average Accuracy	Average F1-Score
XGBoost + TF-IDF	0.7523	0.7322
XGBoost + Word2Vec	0.7660	0.7492
XGBoost + FastText	0.7684	0.7571

As mentioned earlier, the Solo Safari dataset suffers from significant class imbalance, which requires careful handling. The researchers then compared the ADASYN approach, Cost-Sensitive Learning, and the baseline model. All three were applied directly to the model that performed best in the previous experiment, namely the XGBoost-FastText combination. The evaluation results are presented in Table 5, which shows that applying Cost-Sensitive Learning achieved the best results. Controlling the XGBoost algorithm with asymmetric weighting improved its sensitivity to the minority class. These results are the opposite of those for ADASYN, because synthesizing synthetic data in overlapping class areas further undermines the linear boundaries established by FastText.

Table 5. Evaluation of data inequality handling fasttext and XGBoost model

Method	Average Accuracy	Average F1-Score
XGBoost + FastText	0.7684	0.7571
XGBoost + FastText + ADASYN	0.7543	0.7578
XGBoost + FastText + Cost-Sensitive Learning	0.7719	0.7641

The results in Table 5 show that Cost-Sensitive Learning outperformed the other two scenarios in terms of both Accuracy and F1-Score. However, the overall Wilcoxon Signed-Rank test yielded a p-value above the 0.05 threshold, indicating that the observed differences were not statistically significant. The overall test results are shown in Table 6. Despite this, the stability of the F1-Score remains a critical indicator of the model effectiveness in handling minority classes. Overall, the combination of FastText, XGBoost, and Cost-Sensitive Learning remains the best-performing configuration, albeit by a narrow margin compared to other scenarios.

Table 6. Results of the wilcoxon signed-rank test

Scenario	p-value for accuracy	p-value for F1-Score	Notes
Baseline compared to Cost-Sensitive	0.8125	0.4375	Insignificant
Baseline compared to ADASYN	0.1875	1.0000	Insignificant
Cost-Sensitive vs ADASYN	0.1250	0.6250	Insignificant

Evaluation of Prediction Models and Class Restructuring Experiments

The performance of the combination of the FastText and XGBoost models with Cost-Sensitive Learning requires more comprehensive analysis. The performance of this model can be assessed by examining the confusion matrix, which visualizes the distribution of predicted labels. The confusion matrix for this best-performing model scenario is shown in Figure 2.

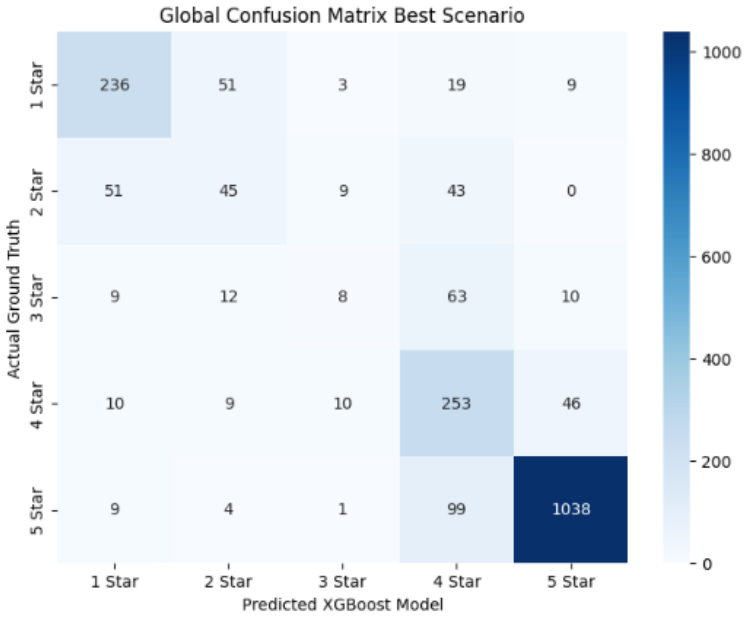

Figure 2. Global confusion matrix (fasttext, cost-sensitive, and XGBoost scenarios)

As revealed in Figure 2, the model indicates good performance in predicting classes with extreme polarity. In this case, the 5-star class gained the highest recall score. Consistent predictions then followed these results in star class 1, which means negative sentiment. This confirms that the weight punishments used in Cost-Sensitive Learning effectively reduce false negatives. Thus, really negative reviews are not easily misclassified or willy-nilly given to other sentiment categories. Although the algorithm can perform optimally on extreme classes, the matrix analysis also reveals a prediction bias that, as shown in the visualization, is concentrated in the middle classes. Like the 3-star class, this class has been shown to have the highest classification error rate. Based on Figure 2, of the 102 actual data points in this 3-star class, the model accurately predicted only 8 reviews. The majority of the data in this class (63 reviews) were incorrectly predicted as 4-star, and the rest were distributed across other classes. This prediction error was caused by the 3-star class being characterized by contradictory sentiment.

To provide a more comprehensive explanation in clarify these arguments, the researchers explored sentiment patterns in the Solo Safari online review dataset for 3-star-rated areas. The observation was conducted to uncover texts that are starkly contradictory within the same label. As seen in the review with ID 295, the visitor expressed positive sentiment about the service, stating, "Secara servis ok, pelayanan & staf ok ramah helpful". However, in the following sentence, the visitor offered sharp, negative criticism of the facilities and animal care, stating, "... Lahan parkir sempit... Binatang masih sedikit... Harimau & singa tampak kurus kurang terawat". A similar pattern of contradiction can also be found in ID 999, where visitors praise facilities such as "... kesannya lbih mewah dan lebih bersih. pelayanan jg ramah" which then clashes with complaints about prices in the statement "... tiket jadi mahal... harganya mehong... Tidak cocok untuk kaum mendang-mending". The nature of 3-star reviews, which blend extreme sentiments, causes the subword embeddings in FastText to process opposing weights simultaneously. This situation causes the algorithm to lose clear decision boundaries, resulting in a vector representation that is quite ambiguous for the 3-star class. A ten-dimensional feature plot of the FastText features that most influence the XGBoost algorithm's decision-making process is shown in Figure 3.

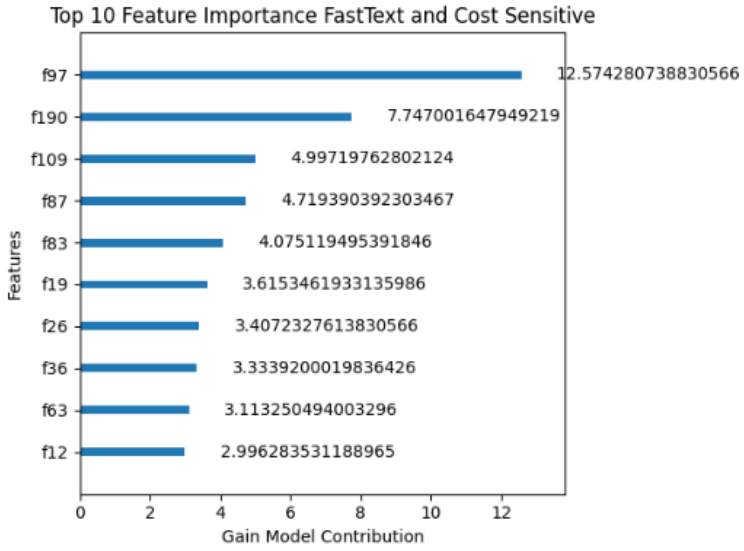

Figure 3. Top ten feature dimensions

The graph in Figure 3 shows only the global feature dimensions (such as f97 and f190) that the XGBoost algorithm relies on most heavily to distinguish among all sentiment classes. The high Gain scores on these dimensions demonstrate that the model is indeed capable of identifying strong discriminating features for predicting review classes. However, even though the model can identify those crucial features, the confusion matrix in Figure 2 shows it still fails to classify the 3-star class. This failure was not caused by a weakness in the XGBoost model, but rather by the ambiguous vectors in the 3-star class. No matter how well the XGBoost model determines decision boundaries at dimensions f97 or f190, the data vectors for 3-star sentiment contradictions will always fall outside these boundaries. Ultimately, the 3-star review is expected to fall into a different category. The feature space's inability to accommodate these contradictory sentiments forms the primary basis for this study's approach to class restructuring and reduction.

This class restructuring experiment was then carried out in stages, beginning with simplifying the dataset into negative, neutral, and positive sentiment classes. The dataset was then simplified by combining 1- and 2-star ratings into the negative sentiment class, 3-star ratings into the neutral sentiment class, and 4- and 5-star ratings into the positive sentiment class. The application of the best prediction model in this study to the restructuring strategy yielded an accuracy of 0.8896 and an F1-score of 0.8825. These prediction results do indeed suggest that restructuring can improve the model's performance. However, these results are not yet optimal, as maintaining the 3-star rating still introduces some noise. Therefore, a subsequent experiment was conducted to eliminate this Class 3 star category from the previous scheme. Therefore, it contains only binary sentiment (positive and negative).

The 3-star class elimination strategy was then applied to improve the performance of the XGBoost prediction model. As it turned out, the model achieved an accuracy and an F1 score of 0.9440. This improvement in the F1 score from 0.8825 to 0.9440 indicates that the 3-star class constitutes noise that can disrupt the algorithm's logic. Thus, grouping classes into a binary sentiment classification and removing the 3-star category can make this model more stable and better at handling sentiment ambiguity.

Business Implications and Service Evaluation

The significance of the predictive model also needs to be further examined in terms of its potential application in Solo Safari's future operations. In this study, the initial dataset had an average score of 4.11. However, after data cleaning, it turned out that the dataset weights had changed due to the elimination of data during the cleaning and annotation processes. The distribution across classes also changed, thereby affecting those average weights. In fact, 5-star ratings declined significantly, ultimately causing the average value in the cleaned dataset to drop to 3.90. The researcher then calculated this average based on the results of the best model in this study to determine the average rating of the predictions for the 5-class scenario. In fact, that average score could rise to 3.94. The difference between the two is only 0.04 points, which actually shows that the model is quite objective and its results are very close to the actual score.

Although it has proven quite accurate from an analytical standpoint, the integration of this predictive model into the Solo Safari management dashboard should initially be positioned as a prototype decision-support system. A binary two-class classification scenario can serve as a practical framework for executives to monitor sentiment trends more quickly. Meanwhile, the operations team can still use the five-category system as an additional reference for monitoring daily complaints. Using these two frameworks can help operational teams recognize the potential for sentiment evaluation bias in the middle class. As a first step, implementing this model still requires direct user evaluation by the Solo Safari management team. Testing in a real-world operational environment must be conducted first to assess the extent to which this model can assist in the daily service evaluation process. Once it has passed a comprehensive feasibility test, this model could be integrated into the service dashboard as an additional tool to help companies respond to visitor feedback more systematically.

4. CONCLUSION

This study can be concluded that the rating prediction model based on XGBoost has been successfully developed in improving the evaluation of Solo Safari visitor satisfaction. With success supported by FastText, Cost-Sensitive Learning, annotation processes, with negative word handling. The main model combining FastText and Cost-Sensitive XGBoost has produced predictions by approaching the average on the original rating, which has a difference value of just 0.04 points. However, the performance is not optimal because there are still many problems with inconsistent online reviews. These challenges cannot be adequately addressed simply by increasing computing power alone, this study proposes that restructuring and class reduction be adopted as the primary strategies. To improve the quality of predictions, neutral reviews with a rating of 3 were not used in this study. In this study, the data used consisted of only two main sentiment categories: positive and negative. This scheme demonstrates excellent model performance, as evidenced by an accuracy and F1-score of 94.40%. Both the 5-class and 2-class strategies can be applied and tested first to the Solo Safari management dashboard with different functions. The 2-class scheme serves as a diagnostic report for executives, and the 5-class scheme can be used by operational teams.

XGBoost does indeed provide a robust modeling foundation. The problem is that the limitations of this model can obstruct a semantic understanding. Future research could explore pre-trained Transformer models such as BERT to address that matter. Also it is suggested to explore tourism-related datasets on a broader scale, beyond Solo Safari. Exploring other datasets can help assess the consistency of this performance across various text patterns. A more comprehensive dictionary of negation terms is also needed to expand the range of negation words. Also, future research could involve linguists with more specialized expertise to reduce noise early on during the annotation stage, thereby maximizing the quality of the feature representations.

ACKNOWLEDGEMENTS

This work was partially supported by the Hibah Penguatan Kapasitas Grup Riset (PKGR-UNS) C, Universitas Sebelas Maret, under contract number 462/UN27.22/PT.01.03/2026. The researchers would like to express their heartfelt gratitude to the university for the support and facilities provided to ensure the steady advancement of this study from start to finish. The researchers also expressed gratitude to three independent annotators for their assistance in the relabeling process, which is specifically because the labeling results can improve the performance of the model.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

The formulation of research ideas, preparation of methods, provision of data, and writing of initial drafts was carried out by **Muhammad Nur Hikmah Ramadhan**. Model validation, support system development, and article review are the responsibility of **Akhmad Syaifuddin**. Meanwhile, **Ristu Saptono** has a role as a supervisor who provides research direction and critically studies the manuscript until the final stage.

DECLARATION OF COMPETING INTERESTS

The author reveals that there is no conflict of interest with related parties, either financially or personally, that can affect the objectivity of the research results shown in this article.

DATA AVAILABILITY

In order to maintain the confidentiality of operational data and ongoing research, the dataset and modeling code are not open to the public. However, access to such data can be obtained through the author of the correspondence by submitting a proper request.

REFERENCES

[1] H. Li, Q. Chen, Z. Zhong, R. Gong, and G. Han, "E-word of mouth sentiment analysis for user behavior studies," *Inf. Process. Manag.*, vol. 59, no. 1, p. 102784, Jan. 2022, doi: 10.1016/j.ipm.2021.102784.

[2] H. Yang and G. Xu, "Tourism attraction selection driven by online tourist reviews: A novel multi-attribute decision making method based on the evidence theory and probabilistic linguistic term sets," *Appl. Soft Comput.*, vol. 177, p. 113243, Jun. 2025, doi: 10.1016/j.asoc.2025.113243.

[3] Y. A. Laghbi and M. Al Dhoayan, "Examining how customers perceive community pharmacies based on Google maps reviews: Multivariable and sentiment analysis," *Explor. Res. Clin. Soc. Pharm.*, vol. 15, p. 100498, Sep. 2024, doi: 10.1016/j.rcsop.2024.100498.

[4] L. Zamani and Khairina, "Kebun Binatang Solo Safari Dibuka Lagi 27 Januari 2023, Harga Tiket di Bawah Rp 50.000," KOMPAS.com. Accessed: Jun. 25, 2026. [Online]. Available: https://regional.kompas.com/read/2022/12/16/160639878/kebun-binatang-solo-safari-dibuka-lagi-27-januari-2023-harga-tiket-di-bawah

[5] R. D. Saputra, J. Peranginangin, and S. Suharto, "Pemanfaatan Instagram Solo Safari sebagai Promosi untuk Meningkatkan Minat Kunjungan," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 3, pp. 7455–7461, Oct. 2025, doi: 10.31004/riggs.v4i3.3106.

[6] N. Kumar and B. R. Hanji, "Aspect-based sentiment score and star rating prediction for travel destination using Multinomial Logistic Regression with fuzzy domain ontology algorithm," *Expert Syst. Appl.*, vol. 240, p. 122493, Apr. 2024, doi: 10.1016/j.eswa.2023.122493.

[7] P. Wang, H. Zhang, X. Yuan, and X. Zhang, "Beyond the stars: Unpacking the impact of score-textual inconsistency of online reviews on hotel performance," *Int. J. Hosp. Manag.*, vol. 130, p. 104271, Sep. 2025, doi: 10.1016/j.ijhm.2025.104271.

[8] A. Almansour, R. Alotaibi, and H. Alharbi, "Text-rating review discrepancy (TRRD): an integrative review and implications for research," *Futur. Bus. J.*, vol. 8, no. 1, p. 3, Feb. 2022, doi: 10.1186/s43093-022-00114-y.

[9] M. T. Uliniansyah *et al.*, "Twitter dataset on public sentiments towards biodiversity policy in Indonesia," *Data Br.*, vol. 52, p. 109890, Feb. 2024, doi: 10.1016/j.dib.2023.109890.

[10] A. Satriamulya and A. Romadhony, "Aspect Extraction on Restaurant Reviews using Domain-Specific Word Embedding," in *2022 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Jul. 2022, pp. 273–277. doi: 10.1109/ICoDSA55874.2022.9862856.

[11] R. B. Madhumala, B. Vineetha, M. R. Shree, R. Sanjesh, and B. R. Charanraj, "A semantic driven model for extraction of text using TF-IDF, SFLA and XGBoost," *Int. J. Inf. Technol.*, vol. 17, no. 6, pp. 3659–3664, Jul. 2025, doi: 10.1007/s41870-025-02549-2.

[12] K. Mokgwatjane and T. Paepae, "An explainable ensemble machine learning approach for multi-domain, multiclass sentiment analysis in Amazon product reviews," *Mach. Learn. with Appl.*, vol. 23, p. 100825, Mar. 2026, doi: 10.1016/j.mlwa.2025.100825.

[13] T. A. Y. Siswa, M. Fattah, M. A. Fahrozi, D. F. Rahman, and F. I. Ramadhani, "Optimizing Sentiment Classification on IKN Development: A Comparative Study of TF-IDF, Word2Vec, and FastText Embeddings," in *MEST 2025*, Basel Switzerland: MDPI, Jun. 2026, p. 20. doi: 10.3390/engproc2026137020.

[14] H. Abdelmotaleb, C. Mcneile, and M. Wojtyś, "A comparative study of word embedding techniques for classification of star ratings," *Expert Syst. Appl.*, vol. 297, p. 129037, Feb. 2026, doi: 10.1016/j.eswa.2025.129037.

[15] S. Ghosal and A. Jain, "Depression and Suicide Risk Detection on Social Media using fastText Embedding and XGBoost Classifier," *Procedia Comput. Sci.*, vol. 218, pp. 1631–1639, 2023, doi: 10.1016/j.procs.2023.01.141.

[16] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.

[17] D. L. Freire, "Cost-Sensitive Algorithms for Text Classification in the Legal Domain: Addressing Imbalanced Law," *Trends Comput. Appl. Math.*, vol. 26, no. 1, p. e01859, Nov. 2025, doi: 10.5540/tcam.2025.026.e01859.

[18] V. Arifin, Y. A. Putri, and R. Wiputra, "A dataset of subjectivity classification in Indonesian ride-hailing app reviews," *Data Br.*, vol. 64, p. 112348, Feb. 2026, doi: 10.1016/j.dib.2025.112348.

[19] H. Wu *et al.*, "PESI: Personalized Explanation recommendation with Sentiment Inconsistency between ratings and reviews," *Knowledge-Based Syst.*, vol. 283, p. 111133, Jan. 2024, doi: 10.1016/j.knosys.2023.111133.

[20] S. Khaled, E. H. Mohamed, and W. Medhat, "Evaluating Large Language Models for Arabic Sentiment Analysis: A Comparative Study Using Retrieval-Augmented Generation," *Procedia Comput. Sci.*, vol. 244, pp. 363–370, 2024, doi: 10.1016/j.procs.2024.10.210.

[21] H. Lee, S. H. Lee, D. Nan, and J. H. Kim, "Predicting User Satisfaction of Mobile Healthcare Services Using Machine Learning: Confronting the COVID-19 Pandemic," *J. Organ. End User Comput.*, vol. 34, no. 6, pp. 1–17, Jun. 2022, doi: 10.4018/JOEUC.300766.

[22] J. C. Puche-Regaliza, I. Marcilla-Lombraña, P. Antón-Maraña, and P. Arranz-Val, "Innovating data-driven tourism reputation management: methodological foundations of the tourism online reputation index (TORI)," *Inf. Technol. Tour.*, vol. 28, no. 1, p. 26, Jun. 2026, doi: 10.1007/s40558-026-00368-0.

[23] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, Nov. 1971,

doi: 10.1037/h0031619.

[24] J. A. Liem, D. T. Djatmiko Utomo, B. Ignasio, and T. W. Cenggoro, "Deep Learning And Machine Learning For Indonesian Online Gambling Website Promotion Detection," *Procedia Comput. Sci.*, vol. 269, pp. 979–992, 2025, doi: 10.1016/j.procs.2025.09.040.

[25] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," in *2018 International Conference on Asian Language Processing (IALP)*, IEEE, Nov. 2018, pp. 226–229. doi: 10.1109/IALP.2018.8629151.

[26] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Nat. Lang. Process. J.*, vol. 2, p. 100003, Mar. 2023, doi: 10.1016/j.nlp.2022.100003.

[27] P. W. Cahyo, U. S. Aesyi, W. A. Setianto, and T. Sulaiman, "A Novel Named Entity Recognition approach of Indonesian fake news using part of speech and BERT model on presidential election," *Int. J. Inf. Manag. Data Insights*, vol. 5, no. 2, p. 100354, Dec. 2025, doi: 10.1016/j.jjimei.2025.100354.

[28] J. Allgaier and R. Pryss, "Cross-Validation Visualized: A Narrative Guide to Advanced Methods," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 1378–1388, Jun. 2024, doi: 10.3390/make6020065.

[29] A. Salleh, M. H. Osman, S. Hassan, M. Y. Said, K. Y. Sharif, and K. T. Wei, "A hybrid model for low-resource language text classification and comparative analysis," *Knowledge-Based Syst.*, vol. 326, p. 114068, Sep. 2025, doi: 10.1016/j.knosys.2025.114068.

[30] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.

[31] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[32] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bull.*, vol. 1, no. 6, p. 80, Dec. 1945, doi: 10.2307/3001968.

[33] S. Ashraf and C. Choi, "XP-STM: A cross-platform sentiment transferability model for negative public sentiment identification and mitigation," *J. Eng. Res.*, vol. 14, no. 2, pp. 1548–1564, Jun. 2026, doi: 10.1016/j.jer.2025.12.005.

.