Scientific Journal of Informatics
Vol. 12, No. 4, November 2025
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Which Features Matter Most? Evaluating Numerical and Textual Features for Helpfulness Classification in Imbalance Dataset using XGBoost

Anindita Putri Kirani[1]*, Rini Anggrainingsih[2], Ristu Saptono[3]
[1,2,3] Department of Informatics, Universitas Sebelas Maret, Indonesia

Abstract.
Purpose: This study aims to develop and realistically evaluate a reliable model for identifying helpful online reviews, particularly in the context of Indonesian-language texts, which are often informal and challenging.
Methods: This study addresses several key challenges in predicting review helpfulness: the relative effectiveness of numerical features from metadata compared with traditional text representations (TF-IDF, FastText) on noisy data; the impact of severe class imbalance; and the limitations of standard validation compared with time-based validation. To address these challenges, we built an XGBoost model and evaluated various feature combinations. A hybrid approach combining SMOTE and *scale_pos_weight* was applied to handle class imbalance, and the best configuration was further assessed using time-based validation to better simulate real-world conditions.
Result: The results show that the model based on numerical features consistently outperformed the text-based model, achieving a peak macro F1-score of 0.7214. Compared to the IndoBERT baseline (F1-score = 0.6400) and the RCNN FastText baseline (F1-score = 0.5317), this indicates that simpler feature-driven models can provide more reliable predictions under noisy review data. Time-based validation further revealed a performance decline of up to 8.06%, confirming the presence of concept drift and highlighting that standard validation tends to yield overly optimistic estimates.
Novelty: The main contribution of this research lies in offering a robust methodology while demonstrating the superiority of metadata-based approaches in this context. By quantifying performance degradation through temporal validation, this study provides a more realistic benchmark for real-world applications and highlights the critical importance of regular model retraining.

Keywords: Review helpfulness, Helpful vote, Time-based evaluation, Imbalanced data handling

Received Sepember 2025 / **Revised** November 2025 / **Accepted** November 2025

INTRODUCTION

Online reviews have become a key source of information for consumers before making purchasing decisions [1] and play an important role in reducing transaction uncertainty [2]. However, the increasing volume of reviews often creates information overload, making it difficult for users to identify truly relevant content [3], [4]. To address this, many platforms provide voting systems that allow users to mark reviews as “helpful” [5]. Unfortunately, these manual systems are slow and prone to bias, highlighting the need for automated models that can proactively identify and display high-quality reviews.

Several prior studies have explored different features for this task [6], [7], [8], but developing a practical review-helpfulness model still faces major challenges that are often overlooked. First, while textual content seems promising, its effectiveness is often hindered in real-world scenarios by the informal nature of the language—full of slang, code-mixing, and typos—a challenge especially pronounced in non-English languages such as Indonesian. Second, as earlier research has shown [9], [10], [11], the severe imbalance between “helpful” and “unhelpful” reviews remains a critical obstacle that can strongly affect model performance. Third, and perhaps most crucial for real-world applications, many studies continue to rely on standard cross-validation, which fails to account for the temporal nature of review data. Since reviews evolve over time, ignoring this aspect can lead to overly optimistic performance metrics that do not reflect long-term model effectiveness. For example, the case of Solo Safari in Surakarta—recently renovated from Jurug Zoo—demonstrates how contextual changes can significantly alter review patterns.

[1]*Corresponding author.
Email addresses: anindyta@student.uns.ac.id (Kirani), rini.anggrainingsih@staff.uns.ac.id (Anggrainingsih), ristu.saptono@staff.uns.ac.id (Saptono)
DOI: 10.15294/sji.v12i4.33443

To address these challenges, this study proposes and evaluates a comprehensive methodology with several key contributions. We systematically compare the performance of numerical metadata-based models with models using traditional text representations (TF-IDF and FastText), aiming to identify the most robust and efficient approach for Indonesia's complex and linguistically diverse context. As a practical contribution, we apply time-based validation to more realistically assess performance degradation caused by concept drift. In addition, we test a hybrid imbalance-handling strategy (SMOTE and scale_pos_weight) to mitigate class bias. Through this approach, the study not only introduces a stronger framework but also underscores the importance of realistic evaluation in developing review-helpfulness models that are truly applicable in practice.

METHODS

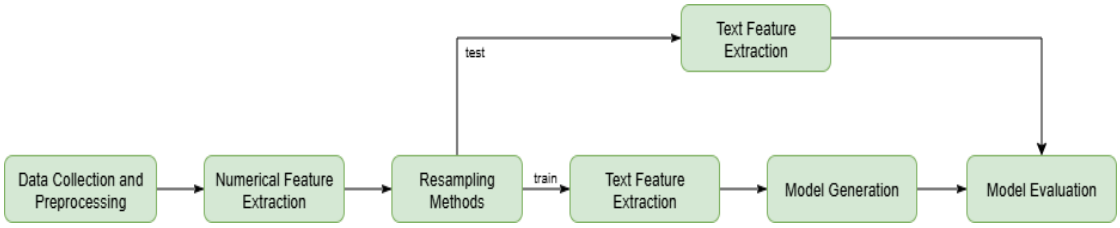

Figure 1. Research Workflow

Data Collection and Preprocessing

The data were collected from the Google Maps platform, which provides reviews of various places worldwide. This study focuses on reviews of a tourist destination in Surakarta, namely Solo Safari. Solo Safari was selected due to its popularity as a tourist attraction that draws both domestic and international visitors [12], as well as its large number of Google Maps reviews, exceeding 17,600 as of January 2025 [13]. The data collection process was carried out using the Google Maps Scraper tool, which applies web scraping techniques. All the reviews analyzed in this study are written in Indonesian. In cases where a review is originally written in another language, the Google Maps platform automatically provides an Indonesian version of the text. Therefore, the dataset used in this research is fully in Indonesian.

Once the data was collected, it went through a cleaning stage to ensure quality and reduce noise that could affect the modeling results. This process involved removing duplicate entries and invalid reviews, such as those with no text or only containing symbols and emojis. Additionally, a cut-off date of May 12, 2025, at 18:02 was set, which was the date of the latest review that had received a like. The cleaned data was automatically labeled based on user interactions. The labeling process was not done manually; instead, it used user engagement on Google Maps as a form of crowdsourced annotation. Each user served as a collective vote, representing the community's perception of a review's usefulness. Reviews with at least one like were labeled as helpful (1), while those with no likes were unhelpful (0). This method, which relies on community-based annotation and user engagement evidence, aligns with other studies that use user reactions to determine review helpfulness [14], [15].

Before proceeding to the next stage, the review data undergoes a preprocessing step to reduce noise and improve data quality. This step includes text cleaning (removing irrelevant characters), text normalization (converting capital letters and standardizing words), and text reduction (removing stopwords and applying stemming). In the normalization stage, non-standard words and slang are standardized using *Kamus Alay* (Colloquial Indonesian Lexicon) [16]. The stemming process is then performed using the Nazief & Adriani algorithm to return words to their root form.

Numerical Feature Extraction

From the review metadata obtained, several features were extracted that were considered relevant to the usefulness of a review. The extracted features are listed in Table 1.

Table 1. List of Numerical Features

Feature	Description
Implicit Features	
Has Numerical Words [17]	Detects the presence of numbers (either digits or Indonesian numeric words). Value = 1 if present; 0 if not.
Has Punctuation Marks [17]	Detects the presence of punctuation marks such as periods, exclamation points, or quotation marks. Value = 1 if present; 0 if not.
Review Length [18]	Number of words in the review text.
Explicit Features	
Answered Any Review Context	Indicates whether the reviewer answered the context question on Google Maps. Value = 1 if they answered at least one; 0 if not.
Stars [19]	Star rating given by the reviewer (scale 1–5).
Is Extreme Rating [20]	Value = 1 if the rating is 1 or 5 (extreme); 0 if the rating is 2, 3, or 4.
Image Count [21]	Number of images attached to the review.
Review Age [22]	Age of the review in days, from the publication date to the experiment cutoff date.
Is Weekend	Value = 1 if the review was written on a weekend; 0 if on a weekday.
Author-Related Features	
Is Local Guide	Value = 1 if the reviewer is a Local Guide; 0 if not.
Reviewer Number of Reviews [23]	The total number of reviews written by the reviewer.

Some features underwent further transformation. For instance, Image Count was categorized based on Google Maps UI/UX insights to form a new feature called *image_category*. Similarly, Review Length and Reviewer Number of Reviews were categorized according to their quantile-based statistical descriptions, resulting in the features *review_length_category* and *reviewer_group*. All transformed data were then encoded using one-hot encoding (OHE) for use in modeling. The research workflow can be seen in figure 1.

Resampling Methods

To evaluate model performance, this study employed stratified 5-fold cross-validation to select the best model. To better simulate real-world conditions, time-based resampling was then applied, specifically time series cross-validation and annual sampling with an expanding window.

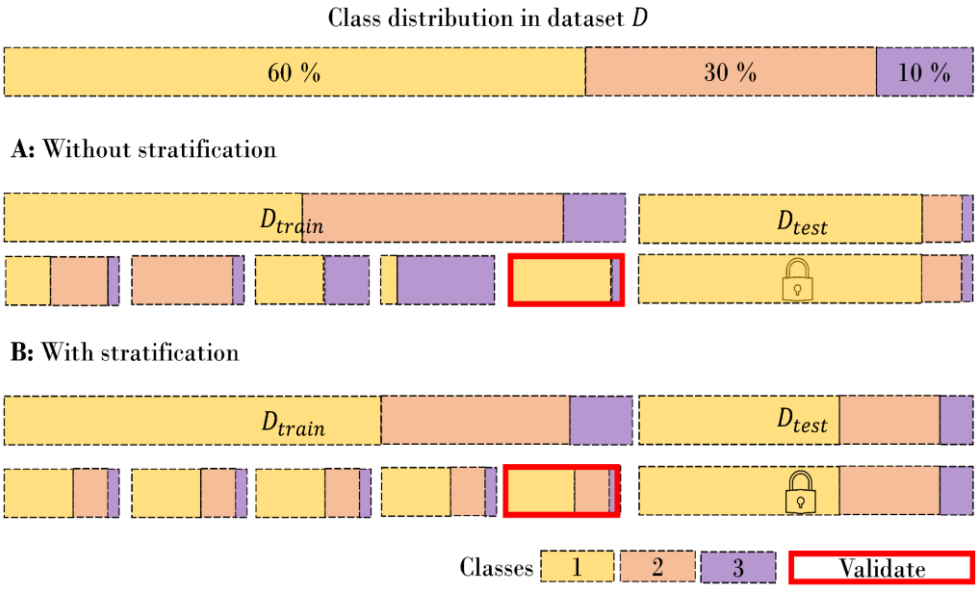

Figure 2. Illustration *K-Fold Cross Validation* with and without *Stratification*

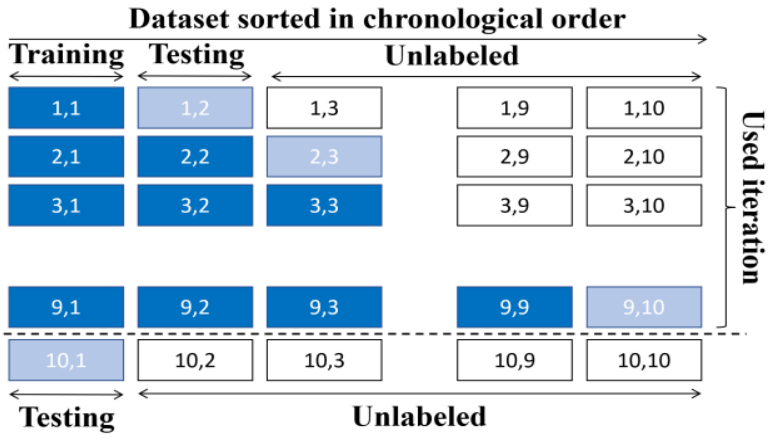

Figure 3. Illustration *Time-Series Cross-Validation* [25] C. *Annual Sampling* [26]

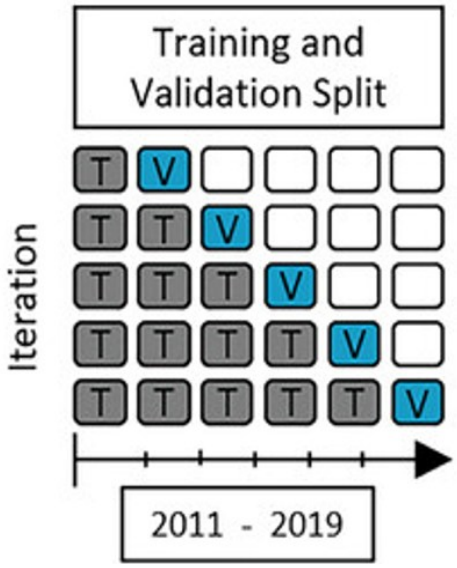

Figure 4. Illustration *Annual Sampling* [26]

Stratified Cross-Validation

In the modeling stage, we used a stratified 5-fold cross-validation method. This technique splits the data into five equal parts, making sure each part has the same proportion of different classes. This is especially useful when the data is not balanced, as it helps the model not focus too much on the most common class and gives a more accurate and fair evaluation of its performance [27]. An illustration of Stratified Cross-Validation can be seen in figure 2.

Time Series Cross-Validation

For time series data, we used a cross-validation method that follows the order of time. Older reviews were used to train the model, and newer ones were used for testing. This approach better mimics real-world situations, where a model learns from past data and is then used to classify new, future data. An illustration of time series cross-validation can be seen in figure 3.

Annual Sampling

To check how the model performs over time and how well it adjusts to yearly changes, we used an annual sampling method with an expanding window. Each year, the training data grows to include more recent information, while the test data is always from the next year. This setup helps us see how well the model can handle changes in how people write reviews, the topics they talk about, and their behavior over time. An illustration of annual sampling can be seen in figure 4.

Text Feature Extraction
Text feature representation was carried out using two approaches: TF-IDF (Term Frequency–Inverse Document Frequency) and FastText pre-trained embeddings. The feature extraction process was applied separately to the training and testing data to avoid data leakage.

TF-IDF was used to represent text based on word frequency adjusted to its uniqueness across documents [28]. On the training data, a fitting process was applied to build the vocabulary and calculate the IDF value. Then, on the testing data, a transformation process was carried out using the TF-IDF model learned from the training data. As a result, words not included in the vocabulary were ignored.

FastText is used to generate dense vector representations of words in text. For both the training and testing data, each word is represented using vectors from a pre-trained FastText model, while the representation of each document is obtained by averaging its word vectors (mean pooling) [29]. If a word is not directly available in the model, FastText can still generate subword-based embeddings, allowing it to provide representation even for previously unseen words or typos.

Model Generation
The classification model was built using the XGBoost algorithm, chosen for its efficiency and high performance [30]. To address the issue of class imbalance, two main approaches were used: XGBoost's scale_pos_weight parameter and resampling with SMOTE [31]. SMOTE resampling was applied under two schemes: 100% and 75%. Optimal hyperparameters for the model were determined through an exhaustive grid search. Furthermore, to provide a comprehensive benchmark for the textual features, the model's performance was compared against two baseline models. The first baseline is an RCNN with FastText embeddings, chosen because this architecture was shown to be a top performer in a similar study by [6]. The second baseline is a fine-tuned IndoBERT, selected because it represents the state-of-the-art (SOTA) for Indonesian text classification tasks and serves as a strong benchmark for any text-based model [4].

Evaluation
To evaluate the performance of classification models in predicting helpful reviews, the evaluation metrics must account for class imbalance in the dataset. Therefore, the macro F1-score was selected as the primary metric. In addition, accuracy was included as a secondary metric.

RESULT AND DISCUSSION
The dataset for this study consists of 6,042 cleaned reviews for the Solo Safari tourist attraction in Surakarta, Indonesia, collected from Google Maps between 2020 and May 12, 2025. The reviews were automatically labeled as helpful (1) if they received at least one "like" and unhelpful (0) otherwise. The raw text underwent a standard preprocessing pipeline including cleaning, case folding, tokenization, normalization with a colloquial Indonesian lexicon, stopword removal, and stemming. Examples of reviews before and after preprocessing can be seen in table 2.

Table 2. Examples of Reviews Before and After Preprocessing

Text Review Before Preprocessing	Text Review After Preprocessing
Saya suka kebun binatang ini! Mereka punya hewan keren, pelayanan bagus. Namun yang kurang saya sukai adalah tempat penampil pertunjukan gajahnya panas. Dan hampir semua barang di sana terlalu mahal, dompet kecil saja harganya 70k. Dan jika kita ingin melihat singa harus membeli tiket premium, kenapa?.	suka kebun binatang mereka punya hewan keren layan bagus yang kurang suka tempat tampil tunjuk gajah panas hampir semua barang sana terlalu mahal dompet kecil harga 70k jika ingin lihat singa beli tiket premium

Two types of features were engineered: numerical and textual. Numerical features, derived from review metadata as detailed in Table 3 and 4, were used to capture structural attributes. Continuous features such as Review Length and Reviewer Number of Reviews were categorized based on statistical quantiles (Q1, Q3), while Image Count was categorized based on UI/UX insights. All resulting categorical features were then one-hot encoded.

Table 3. Summary Statistics of Numerical Features

Features	Alias	Q1 (25%)	Q3 (75%)	Min	Max
Review Length	*review_length*	5	28	1	387
Has Humerical Words	*has_numerical_words*	0	1	0	1
Has Punctuation Marks	*punctuation_marks*	0	1	0	1
Answered Any Review Context	*answeredAnyReviewContext*	0	1	0	1
Stars	*stars*	3	5	1	5
Is Extreme Rating	*is_extrame_rating*	0	1	0	1
Image Count	*image_count*	0	1	0	50
Review Age	*review_age*	394	820	0	1956
Is Weekend	*is_weekend*	0	1	0	1
Is Local Guide	*isLocalGuide*	0	1	0	1
Reviewer Number of Reviews	*reviewerNumberOfReviews*	5	72	1	2038

Table 4. Summary Statistics of Numerical Features

Features	**Alias**	**Q1 (25%)**	**Q3 (75%)**	**Min**	**Max**
Review Length	review_length	5	28	1	387
Has Humerical Words	has_numerical_words	0	1	0	1
Has Punctuation Marks	punctuation_marks	0	1	0	1
Answered Any Review Context	answeredAnyReviewContext	0	1	0	1
Stars	stars	3	5	1	5
Is Extreme Rating	is_extrame_rating	0	1	0	1
Image Count	image_count	0	1	0	50
Review Age	review_age	394	820	0	1956
Is Weekend	is_weekend	0	1	0	1
Is Local Guide	isLocalGuide	0	1	0	1
Reviewer Number of Reviews	reviewerNumberOfReviews	5	72	1	2038

Once the data is prepared, the first step in the modeling process is to split it into training and testing sets using stratified 5-fold cross-validation. After dividing the data, text features were extracted using TF-IDF and FastText. For the TF-IDF representation, sparse vectors were generated, with most values being zero because not all words appear in every document. TF-IDF data representation can be seen in figure 5. Meanwhile, FastText generates dense representations of review texts using subword-based pre-trained embeddings. Each word is converted into a 300-dimensional vector, and mean pooling is then applied to all words in a review to form a sentence vector. FastText data representation can be seen in figure 6.

	00	01	05	06	07	08	0823	09	0an	0er
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 5. Data Representation by TF-IDF

	fasttext_0	fasttext_1	fasttext_2	fasttext_3	fasttext_4
0	0.013794	-0.031403	0.027130	0.092328	0.005345
1	0.015029	-0.019401	0.005181	0.065640	-0.012509
2	0.015922	-0.006192	0.018364	0.074752	-0.009878

Figure 6. Data Representation by FastText

The next step is to build a classification model. The model-building process involves comparing various approaches to handling imbalanced data. Evaluation results are obtained by calculating the average macro F1-score and accuracy across all folds. Model performance details are shown in Table 5 and 6.

Table 5. Model Performance

		Without SMOTE			SMOTE 100%			SMOTE 75%		
	Model	F1 -Score	Accuracy	Time(s)	F1-Score	Accuracy	Time(s)	F1-Score	Accuracy	Time(s)
Baseline	IndoBert	0.6400	0.7825	5851	0.6320	0.7628	16225	0.6385	0.7683	4649
	RCNN FastText	0.5317	0.8059	95	0.5927	0.7289	128	0.5987	0.7436	115
Our Model	Numerik	0.7042	**0.8396**	1.67	0.7183	0.8343	1.48	0.7179	0.8360	1.39
	TF-IDF	0.6114	0.8002	34	0.6494	**0.8045**	58	0.6394	0.8026	37
	FastText	0.5964	**0.8022**	22	0.6493	0.7612	57	**0.6526**	0.7726	36
	Numerik TF-IDF	0.6795	0.8272	36	0.6842	0.8279	61	0.6881	0.8294	67
	Numerik FastText	0.6834	**0.8325**	46	0.6956	0.8231	40	0.6914	0.8284	56

Table 6. Model Performance

	scale_pos_weight			SMOTE 75% + scale_pos_weight		
	F1-Score	Accuracy	Time(s)	F1 -Score	Accuracy	Time(s)
Baseline	-	-	-	-	-	-
	-	-	-	-	-	-
Our Model	0.7037	0.7814	3.05	**0.7214**	0.8282	1.59
	0.6498	0.7519	47	0.6490	0.7956	18
	0.6474	0.7729	113	0.6472	0.7628	74
	0.7078	0.7888	63	0.6993	**0.8348**	61
	0.7078	0.8131	103	0.7089	0.8226	63

Analysis of different feature types shows that numerical features consistently deliver the best performance across all test scenarios. Without any handling of class imbalance, the XGBoost model trained solely on numerical features and achieved a macro F1-score of 0.7042. Its performance further increased to a peak of 0.7214 when applying a combination of SMOTE (75%) and *scale_pos_weight*. This superior performance suggests that the structural attributes or metadata of a reviews such as its length, number of images, and star rating—are stronger and more reliable predictors of helpfulness than textual content in this dataset. In terms of efficiency, numerical features also stand out, with computation times ranging from only 1 to 3 seconds per fold, making them the most pragmatic and efficient approach.

In contrast, the XGBoost model relying solely on textual features (TF-IDF or FastText) demonstrates substantially lower performance, with initial F1-scores of only 0.6114 (TF-IDF) and 0.5964 (FastText). This underperformance can be attributed to the high linguistic complexity and noise in Indonesian review data, a challenge also noted in prior studies [32], [33]. Qualitative analysis reveals that review texts often contain considerable lexical variation, including regional languages (e.g., *suwejuk*, *pokoke jos*), code-mixing (*worth it, recommended*), slang (*mantul*, *kaum mendang-mending*), abbreviations (*sdh/udh = sudah*), acronyms (*HTM*), and typos. Such diversity poses significant difficulties for word-based models like TF-IDF. Although FastText is designed to handle some variation through subword embeddings, its effectiveness appears limited, likely because its pre-training corpus does not sufficiently cover informal or domain-specific vocabulary related to tourism.

However, the analysis shows that this text-based model is highly responsive to data balancing techniques. The application of SMOTE, for example, significantly improved the FastText F1-score to 0.6526—the largest gain among all feature scenarios. This optimized performance placed the XGBoost FastText model on par with the IndoBERT baseline (F1-score = 0.6400). The main advantage of the XGBoost approach lies in its markedly higher computational efficiency, with training requiring only tens of seconds compared to the thousands of seconds needed by IndoBERT. Moreover, the XGBoost FastText model consistently outperformed the RCNN FastText baseline (F1-score = 0.5317), demonstrating that greater architectural complexity does not necessarily lead to better results on noisy data.

The third approach involved combining numerical and textual features. While this hybrid model improved performance relative to purely textual models, its best F1-score (0.7089 for Numerical + FastText) still fell short of the pure numerical model (0.7214). This finding suggests that, within the XGBoost framework, textual features represented by TF-IDF or FastText tend to introduce additional noise rather than meaningful predictive signals. As a result, these noisy textual features slightly disrupt the strong patterns that numerical features alone are able to capture more effectively.

To understand why this numerical model is effective, a feature importance analysis was conducted figure 7, which confirmed that review_length_category, image_category, reviewer_group, and stars are the primary drivers of classification. To further examine how each attribute contributes, a cross-tabulation analysis was performed in table 7.

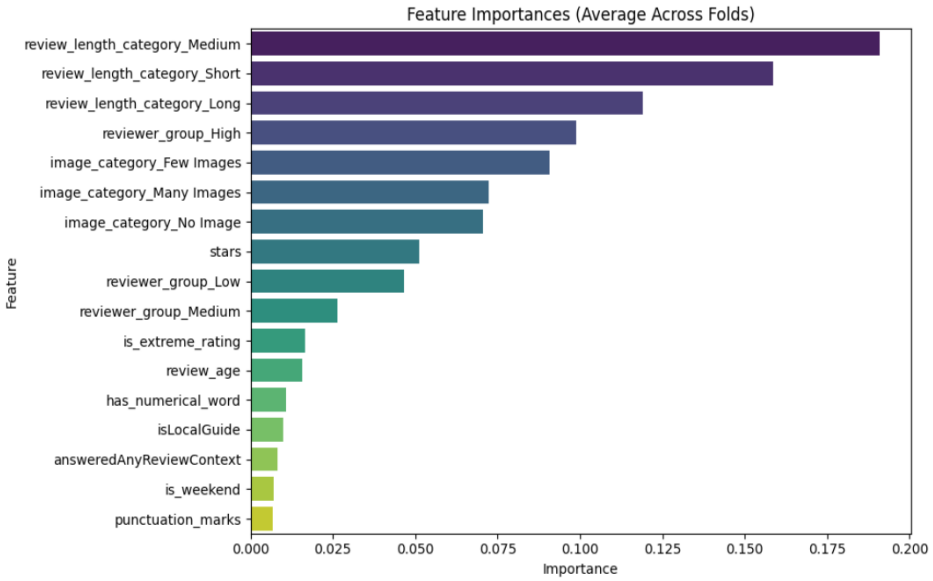

Figure 7. Feature Importance

Table 7. Crosstab Results

Review Length Category			
Features	***Unhelpful* (0)**	***Helpful* (1)**	**Total**
Long	854	596	1.450
Medium	2.420	475	2.895
Short	1.552	145	1.697
Image Category			
Features	***Unhelpful* (0)**	***Helpful* (1)**	**Total**
No Image	3.225	585	3.810
Few Image	1.084	301	1.385
Many Image	517	330	847
Reviewer Group			
Features	***Unhelpful* (0)**	***Helpful* (1)**	**Total**
Low	1.355	271	1.626
Medium	2.344	572	2.916
High	1.127	373	1.500
Stars			
Features	***Unhelpful* (0)**	***Helpful* (1)**	**Total**
1	330	296	626
2	258	120	378
3	645	147	792
4	1.047	246	1293
5	2.546	407	2953

Regarding review length, long reviews produced the highest number of helpful cases (596) compared to the medium and the short with 475 and 145, respectively. This finding aligns with [34], who demonstrated that longer reviews are consistently perceived as more helpful because they provide greater detail. A similar trend was observed in the use of images, where reviews containing many images showed a more balanced

ratio of helpful to unhelpful cases, suggesting that visual evidence enhances perceived quality. This supports [35], who found that visual information complements text and increases review diagnostic. Author reputation also exhibited a clear pattern, with reviewers in the "Medium" and "High" groups contributing the largest share of helpful reviews. The most striking pattern emerged in star ratings. One-star reviews contained nearly equal numbers of helpful (296) and unhelpful (330) cases, while five-star reviews showed the opposite trend, with unhelpful cases (2,546) far exceeding helpful ones (407). This strongly suggests that detailed critical reviews are perceived as more valuable—a finding consistent with prior research [20], which notes that negative reviews are often considered more informative than generically positive ones.

These quantitative patterns provide important context for understanding anomalies and model weaknesses, particularly prediction errors. Cross-tabulation results show that long reviews contributed the highest absolute number of helpful cases (596). The model captured this pattern and tended to classify long reviews as "helpful," especially when accompanied by other strong signals. For example, Review #95 (Table 6) — a very long review with a five-star rating and 28 images—was confidently predicted as helpful. However, its true label was "not helpful," likely due to its overly narrative content. Such cases highlight anomalies of unhelpful long reviews, where the model misinterprets a strong numerical profile as a guarantee of usefulness.

Conversely, in False Negative cases, the model fails to recognize highly informative short reviews, such as Review #4406 (Table 8). This limitation is explained by the dominant training pattern: short reviews are overwhelmingly unhelpful, with only 145 helpful cases compared to 1,552 unhelpful ones. The numerical profile of Review #4406 (short length, moderate four-star rating) fit this majority pattern, leading the model to misclassify it as unhelpful despite its high informational value.

Table 8. Examples of False Positives and False Negatives

ID	Text Review [in Indonesian]	Review Length Category	Image Category	Reviewer Group	Stars	Label	Predicted Label
95	Pertama kali mengajak anak-anak mengunjungi Solo Safari. Ini juga pertama kalinya bagi saya mengunjungi Solo Safari, setelah sebelumnya di tahun 2019 terakhir kali saya mengunjungi Solo Zoo/ Jurug masih dgn konsep yg lama … Semua bentuk pembayaran didalam Solo Safari hanya bisa bertransaksi melalui Qris. Untuk pembelian htm, kita bisa memesan secara online maupun membeli on the spot secara cash maupun cashless.	Long	Many	High	5	0	1
4406	Belum selesai, masih dalam pembangunan	Short	Many	Medium	4	1	0

Overall, the analysis demonstrates that while the model effectively captures general metadata patterns, its main weakness lies in cases where qualitative content quality diverges from quantitative profiles. Specifically, it struggles with low-quality long reviews (leading to False Positives) and high-quality short reviews (leading to False Negatives).

To simulate real-world implementation scenarios, the best-performing model was re-evaluated using two temporal validation schemes, with the results shown in Table 9. Compared to standard cross-validation, performance declined significantly: the Macro F1-score dropped by 6.39% under time-series cross-validation (to 0.6753) and by 8.06% under annual sampling (to 0.6633). This decline underscores a key challenge in deploying machine learning models: concept drift, where data patterns shift over time.

Table 9. Model Performance with *Time-Based Sampling*

	Macro F1	**Accuracy**
Time-Series Cross-Validation	0.6753 ± 0.0276	0.7820 ± 0.0623
Annual Sampling	0.6633 ± 0.0272	0.8097 ± 0.0320

Several factors likely contributed to this decline, including shifts in review topics. The dataset spans periods before and after a major renovation of the tourist site (from Jurug Zoo to Solo Safari). As a result, models trained on older reviews may fail to recognize the context or names of new attractions discussed in more recent reviews. Language has also evolved, with changes in slang and vocabulary trends, while user behavior—such as the frequency of photo uploads—has altered metadata distributions. Temporal validation

therefore offers not only a more realistic benchmark but also emphasizes the need for periodic retraining. Without regular updates, model performance is bound to degrade in dynamic production environments.

Overall, the findings of this study provide several practical lessons that can be applied directly. First, the dominance of simple yet efficient numerical features demonstrates that digital platforms can build fast and accurate review-filtering systems without investing heavily in complex text-processing technologies. This metadata-based approach is particularly well-suited for large-scale applications because it is both cost-effective and practical. Second, the results related to textual features highlight that advanced NLP models are not immediately effective when applied to informal Indonesian-language review data. This limitation arises because reviews often contain slang, abbreviations, code-mixing, or local terms that are not well-represented in standard training corpora. As a result, models require specific adaptations—such as text normalization, dictionary expansion, or additional training with review data—to accurately capture meaning. Finally, the decline in performance observed in temporal validation confirms that real-world prediction systems cannot remain static. To stay relevant and accurate, models must be regularly updated through retraining, which should be treated as an essential part of system maintenance.

CONCLUSION

This study aims to develop and realistically evaluate a reliable model for predicting the usefulness of online reviews, particularly in the context of informal Indonesian language used in tourist destinations. The main findings show that a metadata-based numerical model outperforms models with traditional text representations (TF-IDF and FastText), achieving the highest macro F1-score of 0.7214. The weaker performance of text-based features is mainly due to the challenges of processing informal language, such as slang and code-mixing. The study also highlights that standard cross-validation tends to produce overly optimistic results, whereas temporal validation reveals a performance decline of up to 8.06%, confirming the practical impact of concept drift. The feature importance analysis shows that review length and star rating are the strongest predictors—a finding consistent with [34], who also identified these attributes as key determinants of usefulness. The unique contribution of this study, however, lies in demonstrating that the set of metadata features as a whole is not only important but also superior to textual representations in the context of informal Indonesian-language reviews. Practically, this indicates that digital platforms can build effective and efficient review filtering systems by leveraging relatively stable metadata. It is important to acknowledge a key limitation of this study: the dataset was sourced from a single domain—specifically, one tourist attraction (Solo Safari). This narrow focus limits the generalizability of the findings, as the identified feature patterns may not exhibit the same predictive power in other domains. This limitation suggests several important directions for future research. First, the model should be validated on a larger and more diverse Google Maps dataset that includes other recreational tourist attractions, in order to test the generalizability of the metadata-first approach. A broader dataset also has the potential to yield a more balanced distribution of helpful votes, which could improve model performance. Second, since current textual representations introduce noise, a promising avenue is to explore combining numerical features with more advanced textual representations. Leveraging a fine-tuned contextual language model, such as IndoBERT, for feature engineering could provide richer semantic signals without significantly reducing efficiency, potentially leading to a superior hybrid model.

ACKNOWLEDGEMENT

This work was supported in part by Universitas Sebelas Maret - Hibah Penelitian - Riset Group - Number: 371/UN27.22/PT.01.03/2025.

REFERENCES

[1] N. Aghakhani, O. Oh, D. G. Gregg, and J. Karimi, “Online Review Consistency Matters: An Elaboration Likelihood Model Perspective,” *Information Systems Frontiers*, vol. 23, no. 5, pp. 1287–1301, Sep. 2021, doi: 10.1007/s10796-020-10030-7.

[2] R. Filieri, E. Raguseo, and C. Vitari, “Extremely Negative Ratings and Online Consumer Review Helpfulness: The Moderating Role of Product Quality Signals,” *J Travel Res*, vol. 60, no. 4, pp. 699–717, Apr. 2021, doi: 10.1177/0047287520916785.

[3] H. Hu and A. S. Krishen, “When is enough, enough? Investigating product reviews and information overload from a consumer empowerment perspective,” *J Bus Res*, vol. 100, pp. 27–37, Jul. 2019, doi: 10.1016/j.jbusres.2019.03.011.

[4] M. Bilal and A. A. Almazroi, "Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews," *Electronic Commerce Research*, vol. 23, no. 4, pp. 2737–2757, Dec. 2023, doi: 10.1007/s10660-022-09560-w.
[5] X. (Simon) Hu and Y. Yang, "What makes online reviews helpful in tourism and hospitality? a bare-bones meta-analysis," *Journal of Hospitality Marketing & Management*, vol. 30, no. 2, pp. 139–158, Feb. 2021, doi: 10.1080/19368623.2020.1780178.
[6] A. Alsmadi, S. AlZu'bi, B. Hawashin, M. Al-Ayyoub, and Y. Jararweh, "Employing Deep Learning Methods for Predicting Helpful Reviews," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, IEEE, Apr. 2020, pp. 007–012. doi: 10.1109/ICICS49469.2020.239504.
[7] Y. Wang, J. Wang, T. Yao, and M. Li, "What makes peer review helpfulness evaluation in online review communities? An empirical research based on persuasion effect," *Online Information Review*, vol. 44, no. 6, pp. 1267–1286, Sep. 2020, doi: 10.1108/OIR-07-2018-0216.
[8] Y. Zhou and S. Yang, "Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews," *IEEE Access*, vol. 7, pp. 27769–27780, 2019, doi: 10.1109/ACCESS.2019.2901472.
[9] M. Mahdikhani, "Exploring commonly used terms from online reviews in the fashion field to predict review helpfulness," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100172, Apr. 2023, doi: 10.1016/j.jjimei.2023.100172.
[10] A. Ishtiaq, K. Munir, A. Raza, N. A. Samee, M. M. Jamjoom, and Z. Ullah, "Product Helpfulness Detection With Novel Transformer Based BERT Embedding and Class Probability Features," *IEEE Access*, vol. 12, pp. 55905–55917, 2024, doi: 10.1109/ACCESS.2024.3390605.
[11] K.-K. Lee, H.-H. Lee, S.-J. Cho, and G.-S. Min, "The context-based review recommendation system in e-business platform," *Service Business*, vol. 16, no. 4, pp. 991–1013, Dec. 2022, doi: 10.1007/s11628-022-00502-y.
[12] S. Kurniawan, "Masjid Sheikh Zayed dan Solo Safari Bawa Wisata Solo Pecahkan Rekor di 2024," 2024.
[13] Solo Safari, "Ulasan Pengunjung Solo Safari di Google Maps," 2025.
[14] X. Li, Q. Li, D. Ryu, and J. Kim, "A BERT-based review helpfulness prediction model utilizing consistency of ratings and texts," *Applied Intelligence*, vol. 55, no. 7, p. 455, May 2025, doi: 10.1007/s10489-024-06100-x.
[15] "Sentiment Analysis for Helpful Reviews Prediction," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 7, no. 3, pp. 34–40, Jun. 2018, doi: 10.30534/ijatcse/2018/02732018.
[16] 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.
[17] S. K. Banbhrani, B. Xu, H. Lin, and D. K. Sajnani, "Spider Taylor-ChOA: Optimized Deep Learning Based Sentiment Classification for Review Rating Prediction," *Applied Sciences*, vol. 12, no. 7, p. 3211, Mar. 2022, doi: 10.3390/app12073211.
[18] M. Guha Majumder, S. Dutta Gupta, and J. Paul, "Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis," *J Bus Res*, vol. 150, pp. 147–164, Nov. 2022, doi: 10.1016/j.jbusres.2022.06.012.
[19] R. Venkatesakumar, S. Vijayakumar, S. Riasudeen, S. Madhavan, and B. Rajeswari, "Distribution characteristics of star ratings in online consumer reviews," *Vilakshan - XIMB Journal of Management*, vol. 18, no. 2, pp. 156–170, Jul. 2021, doi: 10.1108/XJM-10-2020-0171.
[20] S. Lee, S. Lee, and H. Baek, "Does the dispersion of online review ratings affect review helpfulness?," *Comput Human Behav*, vol. 117, p. 106670, Apr. 2021, doi: 10.1016/j.chb.2020.106670.
[21] E. Bigne, C. Ruiz, A. Cuenca, C. Perez, and A. Garcia, "What drives the helpfulness of online reviews? A deep learning study of sentiment analysis, pictorial content and reviewer expertise for mature destinations," *Journal of Destination Marketing & Management*, vol. 20, p. 100570, Jun. 2021, doi: 10.1016/j.jdmm.2021.100570.
[22] J. E. Fresneda and D. Gefen, "A semantic measure of online review helpfulness and the importance of message entropy," *Decis Support Syst*, vol. 125, p. 113117, Oct. 2019, doi: 10.1016/j.dss.2019.113117.
[23] M. Siering, J. Muntermann, and B. Rajagopalan, "Explaining and predicting online review helpfulness: The role of content and reviewer-related signals," *Decis Support Syst*, vol. 108, pp. 1–12, Apr. 2018, doi: 10.1016/j.dss.2018.01.004.

[24] 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.
[25] R. Saptono and T. Mine, "Best Approximate Distribution-based Model for Helpful Vote of Customer Review Prediction," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, Oct. 2022, pp. 3427–3434. doi: 10.1109/SMC53654.2022.9945190.
[26] M. Vorndran, A. Schütz, J. Bendix, and B. Thies, "Current Training and Validation Weaknesses in Classification-Based Radiation Fog Nowcast Using Machine Learning Algorithms," *Artificial Intelligence for the Earth Systems*, vol. 1, no. 2, Apr. 2022, doi: 10.1175/AIES-D-21-0006.1.
[27] J. Sadaiyandi, P. Arumugam, A. K. Sangaiah, and C. Zhang, "Stratified Sampling-Based Deep Learning Approach to Increase Prediction Accuracy of Unbalanced Dataset," *Electronics (Basel)*, vol. 12, no. 21, p. 4423, Oct. 2023, doi: 10.3390/electronics12214423.
[28] F. Lan, "Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method," *Advances in Multimedia*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/7923262.
[29] W. Y. Melhem, A. Abdi, and F. Meziane, "Deep Learning Classification of Traffic-Related Tweets: An Advanced Framework Using Deep Learning for Contextual Understanding and Traffic-Related Short Text Classification," *Applied Sciences*, vol. 14, no. 23, p. 11009, Nov. 2024, doi: 10.3390/app142311009.
[30] 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.
[31] S. M. Shrestha and A. Shakya, "A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal," *Procedia Comput Sci*, vol. 215, pp. 652–661, 2022, doi: 10.1016/j.procs.2022.12.067.
[32] A. F. Hidayatullah, "Language tweet characteristics of Indonesian citizens," in *2015 International Conference on Science and Technology (TICST)*, IEEE, Nov. 2015, pp. 397–401. doi: 10.1109/TICST.2015.7369393.
[33] M. A. S. Nasution and E. B. Setiawan, "Enhancing Cyberbullying Detection on Indonesian Twitter: Leveraging FastText for Feature Expansion and Hybrid Approach Applying CNN and BiLSTM," *Revue d'Intelligence Artificielle*, vol. 37, no. 4, pp. 929–936, Aug. 2023, doi: 10.18280/ria.370413.
[34] S. M. Mudambi and D. Schuff, "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.Com1," *MIS Quarterly*, vol. 34, no. 1, pp. 185–200, Mar. 2010, doi: 10.2307/20721420.
[35] Y. Chu, X. Liu, and C. Liu, "The Role of Visual Cues in Online Reviews: How Image Complexity Shapes Review Helpfulness," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 3, p. 181, Jul. 2025, doi: 10.3390/jtaer20030181.