CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 181-193
p-ISSN: 2086-0382; e-ISSN: 2477-3344

BERTopic-Based Multi-Class Topic Classification on Indonesian Shopee E-commerce Reviews Using Ensemble Learning

Kevin Alifviansyah*, Asep Saefuddin, and Septian Rahardiantoro

Department of Statistics and Data Science, School of Data Science, Mathematics and Informatics, IPB University, Indonesia

Abstract

The rapid growth of e-commerce platforms has resulted in a large volume of unstructured user reviews, creating challenges for scalable analysis. This study proposes a multi-class topic classification framework for Indonesian Shopee application reviews by integrating BERTopic-based embedding-driven topic modeling with ensemble learning. A total of 23,956 reviews are analyzed, with BERTopic applied exclusively to 19,167 training reviews to derive eight dominant topic labels, which serve as pseudo-labels for supervised classification using CatBoost and Extra Trees. Model performance is evaluated on a held-out test set under baseline and hybrid resampling settings to address severe class imbalance. The results show that hybrid resampling substantially improves balanced accuracy, particularly for CatBoost, while ROC–AUC remains consistently high, indicating robust class discrimination. Analysis of an unlabeled 2025 dataset, used solely in a deployment-style setting, reveals semantically consistent topic distributions on unseen data. Overall, the findings demonstrate that embedding-based topic modeling combined with ensemble learning provides an effective and scalable solution for multi-class topic classification in highly imbalanced e-commerce review data, with clear separation between training, evaluation, and post-deployment analysis.

Keywords: BERTopic; CatBoost; e-Commerce Reviews; Ensemble Learning; Imbalanced Data; Multi-Class Classification; Topic Modeling.

1. Introduction

The rapid proliferation of e-commerce platforms has resulted in an exponential increase in user-generated textual reviews, positioning them as a high-value data source for both consumer behavior analysis and service quality assessment. These reviews encode multiple latent dimensions of user experience, including product attributes, pricing strategies, logistics performance, and application-level usability. Beyond their role in influencing purchasing decisions, large-scale review corpora provide actionable signals for downstream analytical tasks such as service evaluation, anomaly detection, and operational optimization [1]. However, the inherently unstructured nature, high dimensionality, and scale of textual review data render manual analysis infeasible, necessitating automated and scalable natural language processing (NLP) approaches.

Topic modeling constitutes a core unsupervised NLP technique for uncovering latent semantic structures within large document collections. Classical probabilistic models, most notably

*Corresponding author. E-mail: kevinalifviansyah@apps.ipb.ac.id

Submitted: December 02, 2025 Reviewed: December 15, 2025 Accepted: January 15, 2026
DOI: https://doi.org/10.18860/cauchy.v11i1.37941

Latent Dirichlet Allocation (LDA), rely on bag-of-words assumptions and word co-occurrence statistics, which often limit their ability to capture contextual semantics. Recent advances in representation learning have shifted topic modeling toward embedding-based approaches that leverage contextualized representations produced by transformer-based language models. BERTopic represents a state-of-the-art framework that integrates pretrained sentence embeddings, nonlinear dimensionality reduction, and density-based clustering to generate semantically coherent and interpretable topic representations [2]. Despite its strong performance in exploratory and descriptive analysis, BERTopic remains fundamentally unsupervised and does not directly address the requirements of automated large-scale topic classification.

To bridge this gap, recent research has explored hybrid frameworks that combine unsupervised topic discovery with supervised classification by mapping latent topics to categorical labels. In this setting, topic modeling functions as a weak labeling or pseudo-labeling mechanism that transforms unannotated text corpora into structured training targets for downstream classifiers. However, existing studies predominantly focus on binary or low-cardinality classification tasks and frequently assume relatively balanced class distributions. In contrast, real-world e-commerce review datasets often exhibit highly skewed and long-tailed topic distributions, giving rise to severe multi-class class imbalance, which remains underexplored in the literature.

Severe class imbalance in multi-class settings introduces significant learning challenges, as standard empirical risk minimization tends to favor majority classes, leading to poor minority-class recall and degraded macro-level performance metrics [3]. This issue is further exacerbated in high-dimensional feature spaces derived from contextual embeddings. Ensemble learning methods have therefore gained prominence due to their ability to reduce variance, improve decision boundary stability, and enhance robustness under complex data distributions [4, 5]. In particular, CatBoost, a gradient boosting framework optimized for categorical and high-dimensional inputs, and Extra Trees, a randomized ensemble of decision trees that emphasizes variance reduction through aggressive feature and split randomization, are well-suited for imbalanced multi-class text classification tasks [6, 7].

Motivated by these considerations, this study proposes an end-to-end framework for multi-class topic classification of Indonesian e-commerce reviews obtained from the Shopee application on the Google Play Store. BERTopic is employed to perform embedding-based topic discovery, generating pseudo-labels that represent latent semantic categories. These topic labels are subsequently used as target variables in supervised classification models trained using CatBoost and Extra Trees. Model performance is systematically evaluated under baseline conditions and hybrid resampling strategies to assess classifier robustness, generalization capability, and sensitivity to severe class imbalance.

The contributions of this study are threefold. First, it proposes a hybrid unsupervised–supervised pipeline that integrates BERTopic-based latent topic extraction with ensemble-based multi-class classification for large-scale, imbalanced e-commerce review data. Second, it provides a comparative empirical analysis of CatBoost and Extra Trees, examining their performance characteristics across multiple imbalance-handling configurations. Third, it offers a systematic evaluation of ensemble learning approaches for multi-class topic classification under severe class imbalance, highlighting their feasibility for automated review categorization without dependence on aggressive resampling techniques.

2. Methods

The methodology of this study begins with dataset preparation, followed by a series of text preprocessing steps to clean and normalize the raw e-commerce review data. The cleaned text is then transformed into numerical feature representations using a term-weighting approach to support computational analysis. At this stage, topic modeling with BERTopic is also applied to uncover latent themes within the reviews, producing topic clusters that are subsequently converted into categorical labels and used as the target variable for supervised learning. Following

this stage, the dataset is divided into training and testing subsets for model development. Two ensemble learning algorithms, CatBoost and Extra Trees, are employed to construct classification models for the BERTopic-derived labels. The performance of both models is compared through systematic evaluation, after which the selected model is validated and subsequently utilized for topic classification on new review data. The complete workflow, which integrates topic modeling and ensemble learning for multi-class classification, is illustrated in Fig. 1.

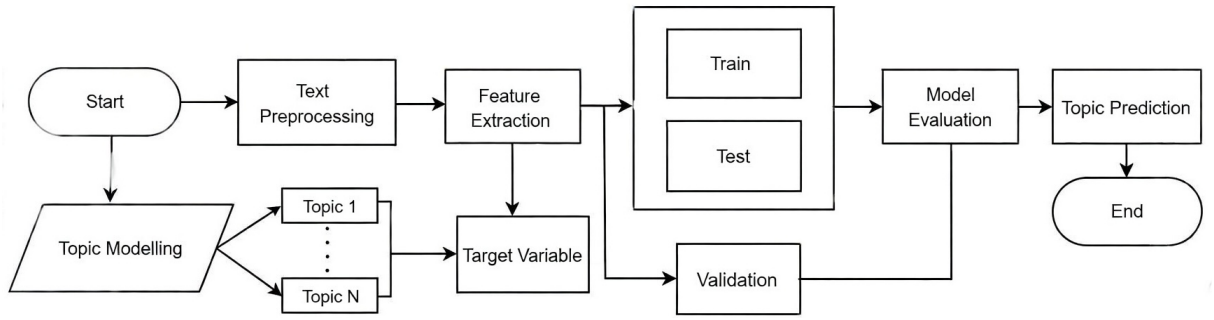

Fig. 1: Research workflow

2.1. Dataset

This study utilizes a corpus of 23,956 Indonesian-language user reviews of the Shopee application collected from the Google Play Store through web scraping. The dataset, stored in CSV format, consists of review text along with associated metadata and was obtained exclusively from publicly accessible sources without involving any personally identifiable information, thereby adhering to ethical standards for secondary data analysis.

After the textual data were consolidated into a structured CSV file, the dataset was split into training and testing sets at the early stage of the analysis. Topic modeling using BERTopic was applied exclusively to the training data to uncover latent thematic structures and to generate topic-based labels serving as the target variable for subsequent supervised classification. All model training and evaluation procedures were conducted based on this train–test separation to ensure methodological validity and to prevent information leakage.

2.2. Text Processing

Text processing aims to normalize and clean raw Indonesian-language e-commerce reviews prior to topic modeling and supervised classification. All preprocessing steps are implemented in a deterministic manner to ensure full reproducibility.

2.2.1. Text Normalization

Text normalization is applied to reduce surface-level variation while preserving the underlying semantic content. In this study, normalization involves converting all characters to lowercase, removing numbers, punctuation marks, emojis, and non-alphabetic characters, normalizing multiple consecutive spaces into a single space, and discarding tokens with fewer than two characters. These procedures effectively reduce noise and sparsity commonly found in informal user-generated review texts.

2.2.2. Tokenization

Tokenization is performed using a word-based tokenizer provided by the *NLTK* library, which segments text based on whitespace and punctuation boundaries. This approach is suitable for Indonesian-language text and remains compatible with subsequent bag-of-words and TF–IDF representations [8].

2.2.3. Stopword Removal

Stopword removal is conducted using the Indonesian stopword list from the *Sastrawi* library. Common functional words such as *dan*, *yang*, *di*, and *ke* are removed to eliminate high-frequency terms that contribute limited semantic value to topic discrimination [8].

2.2.4. Indonesian Stemming

Morphological normalization is performed using the *Sastrawi* stemmer, which implements the Nazief–Adriani algorithm for Indonesian morphology. Stemming is employed to reduce inflected and derived words to their base forms, as it is computationally efficient and has been shown to perform effectively on large-scale Indonesian text corpora [9, 10].

2.3. Topic Modelling with BERTopic

Topic modeling is performed using the BERTopic framework, which integrates contextual sentence embeddings, nonlinear dimensionality reduction, density-based clustering, and class-based TF–IDF representations to generate interpretable topic labels [11].

2.3.1. Text Embedding Model

Document representations are generated using the *IndoBERT-base-p2* model implemented via the *Sentence-Transformers* library. This model is specifically pretrained on large-scale Indonesian corpora and has demonstrated strong performance in downstream semantic representation tasks for the Indonesian language [12, 13]. The resulting sentence embeddings capture contextual semantics beyond surface-level lexical patterns.

2.3.2. Dimensionality Reduction

To project high-dimensional sentence embeddings into a lower-dimensional manifold suitable for clustering, Uniform Manifold Approximation and Projection (UMAP) is applied. In this study, UMAP is configured with 100 nearest neighbors, five output components, a minimum distance of 0.1, and cosine distance as the similarity metric. A fixed random seed is used to ensure reproducibility. This configuration balances local semantic preservation and global structure smoothing, which is beneficial for topic separation in dense embedding spaces [14].

2.3.3. Topic Clustering

Topic clusters are generated using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The clustering model is configured with a minimum cluster size of 50 and a minimum of 15 samples per core point, using Euclidean distance as the clustering metric. The excess of mass (EOM) strategy is employed for cluster selection, with a small cluster selection epsilon of 0.1 to refine cluster boundaries. HDBSCAN is selected for its robustness to variable cluster densities, its ability to identify noise documents, and its suitability for highly imbalanced topic distributions [15].

2.3.4. Topic Representation and c-TF–IDF

To construct interpretable topic representations, BERTopic employs a class-based TF–IDF transformation, which emphasizes words that are distinctive to individual topic clusters rather than globally frequent terms [11]. Term frequency extraction is performed using a *CountVectorizer* configured with Indonesian stopwords obtained from the *Sastrawi* library, bigram inclusion (`ngram_range = (1,2)`), and a minimum document frequency of two.To reduce redundancy among representative topic words and improve interpretability, Maximal Marginal Relevance (MMR) is applied with a diversity parameter of 0.7. This mechanism balances relevance and diversity by penalizing semantically redundant terms within the same topic.

2.3.5. Topic Label Generation

The topic assignments generated by BERTopic are treated as pseudo-labels and subsequently used as the target variable (Y) in the supervised classification stage. Posterior topic probabilities are computed for each document to support stable and reliable topic assignment. As a result of topic reduction, the corpus is organized into eight coherent and interpretable topics, while preserving the underlying density-based structure learned by the model. This reduced topic structure enables controlled multi-class classification and supports robust performance evaluation under severe class imbalance.

2.4. CatBoost

CatBoost is a gradient boosting algorithm based on decision trees that is known for its robustness on high-dimensional feature spaces and its ability to produce well-calibrated class probability estimates [16] . Although CatBoost is originally designed to efficiently handle categorical features through ordered target statistics [17], in this study all input features are numerical vectors derived from TF–IDF representations. Therefore, the advantages related to native categorical feature handling are not exploited.

In this paper, CatBoost is utilized primarily for its strong generalization capability, resistance to overfitting, and stable optimization behavior in multi-class classification with sparse, high-dimensional numerical inputs. The model is trained using a softmax-based multi-class log-loss function, defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log \hat{p}_{i,k}$$

where N denotes the number of samples, K is the number of classes, $y_{i,k}$ is a binary indicator of class membership, and $\hat{p}_{i,k}$ represents the predicted probability for class k.

To address severe class imbalance, two strategies are evaluated. In the baseline setting, CatBoost is trained without resampling while incorporating class weights inversely proportional to class frequencies to reduce bias toward majority classes. In the hybrid setting, the same class-weight configuration is combined with data-level resampling to further improve minority-class recall. Hyperparameters such as tree depth, learning rate, number of iterations, and border count are optimized using randomized search with stratified cross-validation.

2.5. Extra Trees

Extra Trees is an ensemble algorithm that constructs classification models by aggregating outputs from multiple decision trees [18]. This algorithm determines split rules through the random selection of feature subsets and partial random thresholds, introducing additional randomness into the decision-making process [19]. For multiclass classification tasks, Extra Trees is well-suited as it handles multiple categories by extending its splitting criteria to maximize class separation across all target classes [20]. The primary parameters influencing its performance include the number of trees in the ensemble, the number of randomly selected features at each split, and the minimum number of samples required to further divide a node [7]. By incorporating a partially random splitting mechanism, Extra Trees enhances flexibility in decision-making, enabling the model to better capture variations in the data. Additionally, the resampling of data subsets helps mitigate overfitting and improves model accuracy, making it a robust and efficient choice for multiclass classification problems.

2.6. TF-IDF

Term Frequency–Inverse Document Frequency (TF–IDF) is a statistical weighting scheme used to quantify the importance of terms in a document relative to a corpus [21]. In this study,

TF–IDF is used to transform preprocessed review texts into numerical feature vectors suitable for supervised classification.

The TF–IDF formulation adopted follows the standard logarithmic inverse document frequency with additive smoothing[21]. The weight of term j in document i is computed as:

$$w_{i,j} = tf_{i,j} \times \left(\log \left(\frac{N+1}{df_i + 1} \right) + 1 \right)$$

where $tf_{i,j}$ denotes the term frequency of term j in document i, df_j represents the number of documents containing term j, and N is the total number of documents in the corpus. The additive smoothing term is applied to prevent division by zero and to stabilize the weighting of rare terms.

This TF–IDF representation results in high-dimensional and sparse numerical feature vectors, which are subsequently used as inputs for both CatBoost and Extra Trees classifiers.

2.7. Model Performance Evaluation

Model performance is evaluated using balanced accuracy and macro-averaged ROC–AUC to ensure a fair and robust assessment under severe multi-class imbalance.

Let $C \in \mathbb{R}^{K \times K}$ denote the multi-class confusion matrix, where C_{ij} represents the number of samples from the true class i that are predicted as class j. For a given class k, the number of true positives and false negatives are defined as $TP_k = C_{kk}$ and $FN_k = \sum_{j \neq k} C_{kj}$, respectively.

Balanced accuracy is defined as the average class-wise true positive rate, computed as:

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}.$$

For multi-class ROC–AUC evaluation, a one-vs-rest strategy is adopted. The macro-averaged AUC is calculated as the unweighted mean of the class-wise AUC values:

$$\text{AUC}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} \text{AUC}_k.$$

This macro-averaging strategy ensures that each class contributes equally to the final evaluation metrics, thereby preventing dominant classes from biasing the overall performance assessment in imbalanced multi-class settings.

3. Results and Discussion

The dataset used in this study was collected through a web scraping process targeting user reviews of the Shopee e-commerce application available on the Google Play Store. These user-generated reviews constitute a large corpus of naturally occurring textual data that reflects users' experiences, perceptions, and evaluations of the application across various aspects, such as usability, transaction processes, and service performance. As the reviews are generated voluntarily by users in a real-world setting, they provide an ecologically valid source of information for analyzing user opinions in the context of mobile e-commerce platforms. Prior to model development, the collected text data were systematically organized and partitioned into training and testing sets. This data splitting process was conducted at an early stage of the analysis to ensure an unbiased evaluation of model performance and to prevent information leakage between the training and testing datasets, thereby preserving the integrity of the experimental design.

3.1. Pre-processing Result

Text preprocessing was applied separately to the training and testing datasets to standardize the textual input and reduce linguistic noise inherent in user-generated reviews. This separation was

intentionally enforced to prevent information leakage from the testing data into the modeling process. The preprocessing pipeline consisted of lowercasing, tokenization, stopword removal, and stemming. Lowercasing was used to eliminate inconsistencies caused by letter case variations, while tokenization segmented the text into lexical units suitable for computational analysis. Stopword removal eliminated high-frequency terms with limited semantic value, thereby improving feature discriminability. Finally, stemming reduced words to their root forms so that different morphological variants could be represented consistently. Examples illustrating the transformation of raw text at each preprocessing stage are presented in Table 1, demonstrating that the procedure produces a cleaner and more analytically tractable corpus.

Table 1: Text preprocessing results

Step	Result
Original Text	Memudahkan konsumen untuk membeli produk yg di inginkan dengan fitur yang sangat bagus dan mudah di cari, serta sering ada diskon, pengiriman barang cepat.
Normalization	memudahkan konsumen untuk membeli produk yg di inginkan dengan fitur yang sangat bagus dan mudah di cari, serta sering ada diskon, pengiriman barang cepat.
Tokenization	['memudahkan', 'konsumen', 'untuk', 'membeli', 'produk', 'yg', 'di', 'inginkan', 'dengan', 'fitur', 'yang', 'sangat', 'bagus', 'dan', 'mudah', 'di', 'cari', 'serta', 'sering', 'ada', 'diskon', 'pengiriman', 'barang', 'cepat']
Stopword Removal	['memudahkan', 'konsumen', 'membeli', 'produk', 'yg', 'inginkan', 'fitur', 'bagus', 'mudah', 'cari', 'sering', 'diskon', 'pengiriman', 'barang', 'cepat']
Stemming	['mudah', 'konsumen', 'beli', 'produk', 'yg', 'ingin', 'fitur', 'bagus', 'mudah', 'cari', 'sering', 'diskon', 'kirim', 'barang', 'cepat']
Final Text	mudah konsumen beli produk yg ingin fitur bagus mudah cari sering diskon kirim barang cepat

Following the preprocessing stage, the cleaned dataset was partitioned into training and testing subsets to support supervised learning and unbiased model evaluation. This data split was performed prior to any modeling steps to ensure that the testing set remained completely unseen during model development. The allocation of samples between the two subsets is summarized in Table 2. The proportions shown in the table indicate that the split preserves the overall representativeness of the dataset with respect to the generated topic labels, thereby providing a reliable basis for subsequent classification experiments.

Table 2: Train and test proportion

Class	Train Sample	Test Sample
1	10692	3755
2	2647	392
3	1986	188
4	1502	157
5	1282	152
6	796	116
7	182	19
8	80	10

Topic modeling was then conducted exclusively on the training data using the BERTopic framework. Restricting topic extraction to the training set is methodologically necessary to avoid incorporating latent semantic information from the testing data, which could otherwise compromise the validity of downstream performance evaluation. The number of topics was determined by BERTopic's density-based clustering mechanism, which groups documents according to semantic similarity in the embedding space rather than enforcing a predefined number of

clusters. Under the chosen model configuration and data characteristics, this process resulted in the identification of eight coherent and semantically distinct topics, each representing a dominant theme present in the training corpus.

The results of the topic modeling process, summarized in Table 3, show that Generic Reviews is the dominant topic, accounting for more than half of the training data, followed by Product Experience and Delivery Logistics, while the remaining topics App Performance, Payment Issues, User Satisfaction, Ordering Process, and Customer Support are represented by substantially fewer instances. This skewed distribution indicates a pronounced class imbalance among the derived topic labels, which poses a challenge for supervised multiclass classification, as models trained on such data tend to favor majority classes and exhibit reduced sensitivity to underrepresented topics. Therefore, explicit class imbalance handling is necessary to ensure balanced learning and robust performance across all eight identified topics, as discussed in the following subsection.

Table 3: Identified topics from BERTopic

Topic	Topic Label	Topic Interpretation	Count	Frequency (%)
Topic 1	Generic reviews	The app is good, items arrive quickly and match the order. Although there are occasional issues, Shopee promptly processes replacements and refunds. Wishing Shopee continued success.	10692	55.78
Topic 2	Product Experience	Satisfied shopping on Shopee, with cheap prices and good product quality.	2647	13.81
Topic 3	Delivery Logistics	Now all Shopee deliveries use SPX, but delivery is slow and disappointing. Customers want the option to choose shipping services again.	1986	10.36
Topic 4	App Performance	Shopee has become very slow to load, making the app uncomfortable to use and discouraging me from using it.	1502	7.84
Topic 5	Payment Issues	Recently Shopee often has bugs and the app suddenly crashes. This is very frustrating and makes me consider uninstalling if performance does not improve.	1282	6.69
Topic 6	User Satisfaction	Happy shopping on Shopee, always satisfied with the service.	796	4.15
Topic 7	Ordering Process	The app feels quite heavy, especially at night when it is hard to use. Despite the pros and cons, the prices and service are good. I hope Shopee continues to improve.	182	0.95
Topic 8	Customer Support	Helps meet daily needs.	80	0.42

The topic distribution produced by BERTopic reveals a substantial class imbalance, with a single dominant topic accounting for the majority of the training instances. This topic primarily captures reviews that are generic in nature, short, and not explicitly tied to a specific functional aspect of the application. Although such characteristics are often treated as noise in aspect-based analyses, this topic is intentionally retained as a valid target class in the supervised classification setting.

This decision is motivated by both linguistic and methodological considerations. In user-generated reviews, generic comments constitute a prevalent form of expression and convey meaningful contextual information related to overall user perception, satisfaction, and service reliability. Removing this topic from the training data would eliminate a large portion of naturally occurring discourse and distort the semantic distribution learned during the topic modeling stage.

Furthermore, preserving all BERTopic-derived labels, including high-frequency and low-specificity classes, enables the classifier to learn decision boundaries that more accurately reflect the underlying structure of the original data. Treating the generic topic as noise and excluding it from model training would introduce selection bias, as the classifier would be trained on an

artificially filtered subset of the data. Therefore, this approach prioritizes semantic completeness and contextual fidelity, while the impact of class imbalance is addressed through dedicated imbalance-handling strategies during the model training phase.

3.2. Addressing Class Imbalance

The training dataset employed for multiclass topic classification, derived from the BERTopic labeling process, exhibits a pronounced class imbalance. This imbalance is characterized by the dominance of a single majority class alongside the limited representation of several minority topic categories. Such a distribution can adversely influence supervised learning algorithms by biasing the training process toward the majority class and constraining the model's ability to adequately learn discriminative patterns associated with less frequent topics.

To mitigate this issue, a hybrid resampling strategy was applied exclusively to the training data in order to prevent information leakage. In this approach, undersampling was conducted on the majority class to reduce its dominance, while oversampling was applied to the remaining minority classes to enhance their representation. This combined resampling procedure yields a more balanced class distribution across all topic categories while maintaining informative samples from the original data. Furthermore, hyperparameter optimization was performed using RandomizedSearchCV with five-fold stratified cross-validation to ensure robust model selection under imbalanced conditions. The effectiveness of this strategy was subsequently assessed using CatBoost and Extra Trees classifiers, both of which are well suited for addressing complex multiclass classification problems under class imbalance.

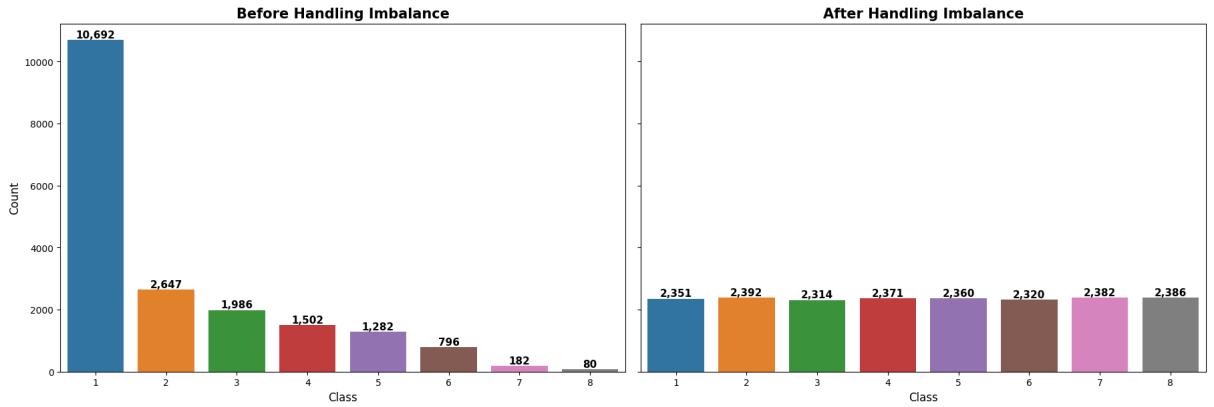

Fig. 2: Class distribution before and after hybrid-resampling

Fig. 2 illustrates the class distribution before and after the application of the hybrid resampling strategy on the training data. Prior to resampling, the distribution is highly skewed, with one dominant class containing a substantially larger number of instances than the remaining classes. After the imbalance handling procedure is applied, the class distribution becomes markedly more uniform across all topic categories. This visual evidence indicates that the combined resampling approach effectively reduces class dominance and enhances the representation of minority classes, thereby providing a more suitable foundation for training supervised multiclass classification models. Based on this balanced data configuration, the optimal model parameters were subsequently determined, as presented in Table 4.

Table 4: Best hyperparameter

	CatBoost	**Extra Trees**
Hyperparameter	{'learning_rate': 0.1, 'l2_leaf_reg': 6, 'iterations': 800, 'depth': 6}	{'n_estimators': 800, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 10}

3.3. Model Evaluation

Table 4 presents the optimal hyperparameter settings obtained from the best evaluation results after class imbalance handling, representing the parameter configurations that achieved the highest performance in the imbalanced multiclass setting.

For the CatBoost classifier, the optimal configuration consists of a learning rate of 0.1, *l2_leaf_reg* of 6, 800 iterations, and a tree depth of 6. This combination balances learning efficiency and regularization, where moderate tree depth and explicit L2 regularization help control model complexity and ensure stable convergence. In contrast, the Extra Trees classifier achieves optimal performance with 800 estimators, *min_samples_split* of 5, *min_samples_leaf* of 2, *max_features* set to *sqrt*, and a maximum tree depth of 10. The large ensemble size and randomized feature selection enhance robustness, while constraints on node splitting and leaf size mitigate overfitting and improve generalization.

Overall, the selected hyperparameters emphasize controlled model complexity, effective regularization, and ensemble stability, which are critical for reliable performance in imbalanced multiclass classification tasks. Table 5 further summarizes the comparative performance of CatBoost and Extra Trees across multiple evaluation metrics, enabling a concise assessment of their relative effectiveness under the selected configurations.

Table 5: Performance comparison model

Model	Balanced accuracy	AUC
Catboost	0.417	0.909
Catboost + hybrid resampling	0.703	0.907
Extra Trees	0.668	0.914
Extra Trees + hybrid resampling	0.674	0.912

Table 5 presents a performance comparison of the CatBoost and Extra Trees classifiers before and after the application of the hybrid resampling strategy, evaluated using balanced accuracy and AUC.

For the CatBoost model, the application of hybrid resampling leads to a substantial improvement in balanced accuracy, increasing from 0.417 to 0.703. This gain indicates a markedly improved ability of the model to handle class imbalance by achieving more balanced performance across all classes. The AUC value remains consistently high, with only a marginal decrease from 0.909 to 0.907, suggesting that the model's overall discriminative capability is largely preserved after resampling.

Similarly, the Extra Trees classifier exhibits strong performance in terms of balanced accuracy, achieving a value of 0.668 without resampling and 0.674 after the hybrid resampling strategy is applied. Although the improvement in balanced accuracy is more modest compared to CatBoost, the results indicate increased robustness under imbalanced conditions. The AUC values remain stable, decreasing slightly from 0.914 to 0.912, which implies minimal impact on the model's ability to distinguish between classes.

Overall, the results demonstrate that the hybrid resampling strategy is particularly effective in improving balanced accuracy, especially for CatBoost, while maintaining relatively stable AUC values for both models. This suggests that hybrid resampling enhances sensitivity to minority classes without substantially degrading overall classification discrimination, reinforcing its suitability for imbalanced multiclass classification tasks.

3.4. Topic Classification

The topic classification process is conducted to evaluate the generalization ability of the best-performing CatBoost model when applied to newly collected validation data from 2025. This analysis serves to assess the model's robustness and adaptability in identifying topical patterns from unseen textual inputs that have undergone standardized text preprocessing and TF-IDF vectorization.

Table 6: Topic Classification Result

Topic	Topic Classification Label	Topic Frequency (%)
Topic 1	Generic reviews	59.59
Topic 2	Product Experience	13.47
Topic 3	Delivery Logistics	8.81
Topic 4	App Performance	6.22
Topic 5	Payment Issues	5.70
Topic 6	User Satisfaction	3.63
Topic 7	Ordering Process	2.07
Topic 8	Customer Support	0.52

This section presents a descriptive analysis of topic classification results obtained from user reviews collected in 2025. As the 2025 dataset does not include ground-truth topic labels, the analysis is positioned as a deployment-oriented examination of how the trained model organizes and characterizes incoming, unlabeled data in a real-world application setting, rather than as an external validation of predictive performance.

The topic distribution shown in Table 6 indicates that *Generic reviews* constitute the largest proportion of the 2025 data, accounting for more than half of all classified instances. This pattern suggests that a substantial share of user feedback during this period is expressed in broad or non-specific terms. Topics associated with more concrete aspects of the user experience, such as *Product Experience*, *Delivery Logistics*, and *App Performance*, appear with moderate frequencies, reflecting recurring but secondary areas of user attention. In contrast, topics related to specific operational or service-related issues, including *Payment Issues*, *Ordering Process*, and *Customer Support*, represent smaller proportions of the overall dataset.

From an application perspective, these results provide a high-level overview of the thematic structure present in the 2025 review corpus as inferred by the deployed model. The relative frequencies of topics may be used to support exploratory analysis, trend monitoring, or prioritization of areas for further qualitative investigation. However, because the classifications are generated without reference to labeled ground truth, the findings should be interpreted as indicative patterns rather than as definitive measurements of model accuracy or temporal generalization.

To substantiate claims related to robustness or external validity across time, future work would require the annotation of a representative subset of the 2025 reviews using a well-defined labeling protocol and appropriate quality assurance procedures. Performance metrics computed on such a labeled subset would enable a more rigorous assessment of model behavior on data drawn from a different temporal context.

4. Conclusion

This study presents an integrated framework for multi-class topic classification of Indonesian e-commerce reviews by combining BERTopic-based embedding-driven topic modeling with ensemble learning. By treating BERTopic-generated topics as pseudo-labels, the proposed approach enables supervised classification without manual annotation while maintaining semantic interpretability of the topic structure.

Experimental results show that severe class imbalance substantially affects classification performance in multi-class settings. The application of a hybrid resampling strategy significantly improves balanced accuracy, particularly for the CatBoost classifier, indicating enhanced sensitivity to minority classes while preserving overall discriminative capability as reflected by stable ROC–AUC values. Compared to Extra Trees, CatBoost demonstrates more consistent and robust performance across imbalance-handling configurations.

Evaluation on an independent dataset collected in 2025 further indicates that the trained model produces semantically coherent topic distributions on unseen data, supporting its applicability in

real-world deployment scenarios. Although the absence of ground-truth labels limits quantitative validation, the results suggest that the proposed framework provides an effective, scalable, and interpretable solution for multi-class topic classification of highly imbalanced e-commerce review data.

CRediT Authorship Contribution Statement

Kevin Alifviansyah: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Visualization, Writing–Original Draft Preparation, Writing–Review & Editing. **Asep Saefuddin:** Supervision, Methodology, Validation, Formal Analysis, Writing–Review & Editing. **Septian Rahardiantoro:** Methodology, Validation, Resources, Writing–Review & Editing, Visualization.

Declaration of Generative AI and AI-assisted technologies

The author used the generative AI tool ChatGPT (version 4, OpenAI) to support idea development and exploration in code writing and LaTeX formatting. In addition, Grammarly was used to assist with grammar checking and refinement of the English text. All outputs generated by these tools were manually reviewed and adjusted to ensure accuracy, originality, and academic integrity.

Declaration of Competing Interest

The authors declare no competing interests.

Data Availability

The data and python scripts used in this study were generated and executed in Google Colab. The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

[1] I. P. Nuralam, N. Yudiono, M. R. A. Fahmi, E. S. Yuliaji, and T. Hidayat. "Perceived ease of use, perceived usefulness, and customer satisfaction as driving factors on repurchase intention: The perspective of the e-commerce market in Indonesia". In: *Cogent Business & Management* 11.1 (2024). DOI: 10.1080/23311975.2024.2413376.

[2] M. Mishra. "A holistic review of customer experience research: Topic modelling using BERTopic". In: *Marketing Intelligence & Planning* (2024). DOI: 10.1108/MIP-09-2023-0457.

[3] S. Das, S. S. Mullick, and I. Zelinka. "On supervised class-imbalanced learning: An updated perspective and some key challenges". In: *IEEE Transactions on Artificial Intelligence* 3.6 (2022), pp. 973–993. DOI: 10.1109/TAI.2022.3160658.

[4] Lukmanul Hakim, Bagus Sartono, and Asep Saefuddin. "Bagging Based Ensemble Classification Method on Imbalance Datasets". In: 2017. https://api.semanticscholar.org/CorpusID:212484809.

[5] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour. "Boosting methods for multi-class imbalanced data classification: An experimental review". In: *Journal of Big Data* 7 (2020), pp. 1–47. DOI: 10.1186/s40537-020-00349-y.

[6] A. N. A. Aldania, A. M. Soleh, and K. A. Notodiputro. "A comparative study of CatBoost and double random forest for multi-class classification". In: *Jurnal RESTI* 7.1 (2023), pp. 129–137. DOI: 10.30598/barekengvol19iss1pp227-236.

[7] A. Sharaff and H. Gupta. "Extra-tree classifier with metaheuristics approach for email classification". In: *Advances in Computer Communication and Computational Sciences*. Springer, 2019, pp. 189–197. DOI: 10.1007/978-981-13-6861-5_17.

[8] Slamet Riyanto, Sukaesih Sitanggang Imas, Taufik Djatna, and Tika Dewi Atikah. "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification". In: *International Journal of Advanced Computer Science and Applications* 14.6 (2023). DOI: 10.14569/IJACSA.2023.01406116.

[9] Bambang Nazief and Mirna Adriani. "Confix Stripping Approach in Indonesian Stemming Algorithm". In: *Proceedings of the Workshop on Computational Linguistics* (1996), pp. 1–13. https://dl.acm.org/doi/10.1145/1316457.1316459.

[10] Indra, Edi Winarko, and Reza Pulungan. "Trending topics detection of Indonesian tweets using BN-grams and Doc-p". In: *J. King Saud Univ. Comput. Inf. Sci.* 31.2 (Apr. 2019), pp. 266–274. DOI: 10.1016/j.jksuci.2018.01.005.

[11] Maarten Grootendorst. "BERTopic: Neural Topic Modeling with a Class-Based TF–IDF Procedure". In: *arXiv preprint arXiv:2203.05794* (2022). DOI: DOI:10.48550/arXiv.2203.05794.

[12] Bryan Wilie, Kevin Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, and Pascale Fung. "IndoBenchmark: Benchmarking Natural Language Processing Tasks for Indonesian". In: *Proceedings of the 28th International Conference on Computational Linguistics* (2020), pp. 843–857. DOI: DOI:10.48550/arXiv.2009.05387.

[13] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (2019). DOI: DOI:10.18653/v1/D19-1410.

[14] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2018). DOI: DOI:10.48550/arXiv.1802.03426.

[15] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical Density Based Clustering". In: *Journal of Open Source Software* 2.11 (2017), p. 205. DOI: DOI:10.21105/joss.00205.

[16] Liudmila Ostroumova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. "CatBoost: unbiased boosting with categorical features". In: *Neural Information Processing Systems*. 2017. https://api.semanticscholar.org/CorpusID:5044218.

[17] John T. Hancock and Taghi M. Khoshgoftaar. "CatBoost for big data: an interdisciplinary review". In: *Journal of Big Data* 7 (2020). DOI: DOI:10.1186/s40537-020-00369-8.

[18] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine Learning* 63 (2006), pp. 3–42. DOI: DOI:10.1007/s10994-006-6226-1.

[19] Budi Padmaja, Vicky Prasa, and K. V. N. Sunitha. "A Novel Random Split Point Procedure Using Extremely Randomized (Extra) Trees Ensemble Method for Human Activity Recognition". In: *EAI Endorsed Trans. Pervasive Health Technol.* 6 (2020), e5. https://api.semanticscholar.org/CorpusID:219545647.

[20] Chalvina Izumi and Nidya Sari Rahmawati. "Handling Multiclass Imbalance in Diabetes, Cancer, and Pneumonia Classification Using NR-Clustering SMOTE". In: *IJACI : International Journal of Advanced Computing and Informatics* (2025). https://api.semanticscholar.org/CorpusID:282367647.

[21] Juan Enrique Ramos. "Using TF-IDF to Determine Word Relevance in Document Queries". In: 2003. https://api.semanticscholar.org/CorpusID:14638345.