MJI
Media Jurnal Informatika
Vol. 18, No. 1, June 2026
p-issn : 2088-2114, e-issn : 2477-2542

Hybrid IndoBERT and Support Vector Machine for Multi-class Emotion Classification of Indonesian Tourism Reviews

Firas Atqiya [a,1,*], Afrida Helen [a,2], Muhammad Rizqi Sholahuddin [b,3]

[a] Department of Computer Science, Math and Science Faculty, Universitas Padjadjaran, Bandung 45363, Indonesia
[b] Computer Engineering and Informatics Department, Politeknik Negeri Bandung, Kab. Bandung Barat 40559, Indonesia
[1] firas.atqiya@unpad.ac.id*; [2] helen@unpad.ac.id; [3] muhammad.rizqi@polban.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 2026-06-08
Revised 2026-06-22
Accepted 2026-06-22

Keywords
emotion classification
IndoBERT
support vector machine
SMOTE
class imbalance

ABSTRACT

Online reviews hold emotional nuances that binary sentiment analysis cannot adequately capture for targeted tourism management. Indonesian reviews pose additional computational challenges due to informal language, Sundanese vernacular, and severe class imbalance. Objective: This study develops a hybrid classification framework using IndoBERT as a frozen feature extractor and a Support Vector Machine (SVM) across five emotional classes. It investigates integrating Principal Component Analysis (PCA) and SMOTE within a strict cross-validation pipeline to mitigate extreme minority class scarcity while preventing data leakage. The duplicate-free dataset comprises 446 manually annotated reviews from agro-tourism destinations in Rancakalong. Annotations followed Ekman's emotions plus a neutral category, cross-validated by a Large Language Model (Cohen's Kappa = 0.7475). To satisfy oversampling constraints, three extreme minority classes (fear, surprise, disgust) were consolidated into an 'OTHER' class. Three configurations were evaluated via 5-Fold Stratified Cross-Validation: TF-IDF + SVM (M1 baseline), IndoBERT + SVM (M2), and IndoBERT + PCA + SMOTE + SVM (M3), utilizing Macro F1 as the primary metric. Results: The M1 baseline yielded a Macro F1 of 0.3920. By capturing contextual semantics, M2 improved accuracy to 0.7131 and Macro F1 to 0.4133. The proposed M3 architecture achieved the highest Macro F1 (0.4321), demonstrating that combining dimensionality reduction and oversampling strengthens minority class decision boundaries. However, erratic performance on the synthetic 'OTHER' class confirms that merging distinct emotions disrupts cohesive semantic signatures. Integrating frozen IndoBERT embeddings with PCA and SMOTE within a cross-validated SVM architecture significantly outperforms traditional baseline models on highly imbalanced, low-resource Indonesian text data. This study contributes an empirically validated emotion corpus and establishes a foundational, data-driven behavioral modeling framework to guide targeted managerial interventions in local agro-tourism.

1. Introduction

Online reviews have become a common record of visitor experience. On platforms such as Google Maps, visitors describe what they liked, what disappointed them, and how a place made them feel. For tourism managers, this text is a low-cost source of feedback. The problem is how to read it at scale. Conventional sentiment [1] analysis sorts text into positive, negative, or neutral. That view is too coarse for targeted tourism management, where a one-star review about a steep, slippery path expresses fear, while a one-star review about garbage expresses disgust, and the two call for different managerial responses. Emotion classification offers a finer-grained perspective. Ekman's theory of basic emotions (joy, sadness, anger, disgust, fear, surprise) [2] provides a robust, compact taxonomy. While other foundational models exist, such as Plutchik's wheel [3], Ekman's set maps cleanly onto short review texts. Adding a neutral class

effectively covers purely descriptive text such as opening hours or ticket prices.

Applying this taxonomy to Indonesian tourism reviews poses significant computational challenges. The language is highly informal and full of slang, reviews frequently mix Indonesian with local Sundanese vernacular, and the distribution of emotions is inherently skewed: most reviews are positive, while negative classes appear only a handful of times. Traditional machine learning architectures utilizing TF-IDF features coupled with a Support Vector Machine (SVM) [4] remain a strong, fast baseline for text classification. However, frequency-based models inherently fail to capture the deep semantic context of sentences.

Recent advancements in Natural Language Processing (NLP) over the past five years have established Transformer-based architectures [5] as the gold standard. For Indonesian text, pretrained models such as IndoBERT [6], [7] and IndoBERTweet [8] capture intricate word order and contextual nuances. Recent empirical studies explicitly highlight the superiority of IndoBERT over traditional methods for analyzing Indonesian tourism reviews and public sentiment [9], [10]. However, full fine-tuning of these massive models typically requires enormous labeled datasets and significant computational overhead, which are often unavailable in localized, low-resource agro-tourism studies. A highly efficient alternative is to utilize the IndoBERT model as a frozen feature extractor, passing its dense semantic representations to a traditional SVM classifier. Recent literature confirms that this hybrid approach achieves near state-of-the-art performance while drastically reducing computational costs [11].

Despite these architectural improvements, handling extreme class imbalance remains a critical bottleneck. While the Synthetic Minority Over-sampling Technique (SMOTE) [12] is the foundational algorithm for generating synthetic minority instances, applying it directly to high-dimensional embeddings (e.g., the 768 dimensions of IndoBERT) frequently generates noisy, out-of-manifold data. Recent methodologies in Indonesian text classification emphasize that applying dimensionality reduction, such as Principal Component Analysis (PCA), prior to oversampling is strictly required to stabilize the feature space and prevent data leakage during cross-validation [13].

Current Indonesian sentiment and emotion studies often compare several machine learning models on a single corpus [10], [14], [15]. However, there is a distinct gap in the literature regarding a rigorous pipeline that combines an original Indonesian agro-tourism dataset, a frozen IndoBERT-SVM architecture, and an explicit evaluation of PCA and SMOTE for extreme class imbalance. This study addresses that gap by formulating three main research objectives. First, it evaluates the performance of a hybrid IndoBERT and SVM model on a five-class emotion classification of Indonesian tourism reviews compared with an SVM and TF-IDF baseline. Second, it measures how PCA and SMOTE affect predictive performance when the class imbalance is extreme. Third, it determines the spatial emotion profile of each destination to provide actionable insights for site managers.

Consequently, the main contribution of this research is establishing a preliminary behavioral modeling framework for agro-tourism. By focusing strictly on emotion model development rather than the immediate creation of a full information system, this study provides a data-driven foundation that could, in principle, support site managers in capturing visitor sentiments. If scaled properly, this foundational layer aligns with the broader objectives of sustainable local tourism, decent work, and sustainable communities (SDG 8 and SDG 11).

2. Method

2.1 Type and Approach of Research

This is a quantitative, experimental study. We treat emotion classification as a supervised learning problem and compare model variants under a fixed evaluation protocol. The experimental design is appropriate because the research questions ask about measured performance differences between models and about the measured effect of an oversampling technique.

2.2 Object and Scope of Research

The object of study is short Indonesian tourism review text and the emotion it expresses. The domain is tourism in Rancakalong, Sumedang Regency, West Java. The scope is limited to seven destinations and to five final emotion classes. The study does not cover aspect-level emotion or multi-label emotion, and it does not fine-tune the language model.

2.3 Data Collection Techniques

We scraped public reviews from Google Maps for seven destinations: Batu Alam, Geotheater Rancakalong, Wisata Alam Paniisan, Panenjoan Pasir Biru, Coffee Buhoen Nagarawangi, Curug Pasirwangi, and Situ Lembang Rancakalong. Inclusion criteria were that a review is written in Indonesian or mixed Sundanese, contains descriptive content rather than a bare rating, and relates to the visitor experience. The final corpus has 457 reviews with the fields Id, Place, Review, and Label.

Annotation used a two-stage process. In the first stage a researcher labelled all 457 reviews against a seven-class taxonomy (joy, sadness, anger, disgust, fear, surprise, neutral) using written mapping rules based on Ekman plus neutral. The rules covered negation (for example, "tidak mahal" maps to joy, not anger) and emotion dominance (a review that praises a place but ends with a sharp complaint about it being abandoned shifts to sadness). In the second stage a large language model acted as an independent validator. Instruction-tuned large language models [16], [17] can follow written labelling guidelines, and recent work reports that they can match or exceed crowd annotators on text-annotation tasks [18], which is why one was used here. It received the same definitions and returned, for each review, a predicted label, a status of agree or disagree, and a short linguistic justification. Disagreements were re-examined by the researcher, and the final ground truth is the reconciled label set in Dataset_Wisata_RancakalongSumedang-humanv3.csv. Agreement between the human and the model was measured with Cohen's Kappa.

2.4 Tools and Materials Used

We ran all experiments in Google Colab on a Tesla T4 GPU with PyTorch 2.11.0. The language model is indobenchmark/indobert-base-p1 loaded through the Hugging Face Transformers library [19]. Classifiers, metrics, and the data split come from scikit-learn [20]. Oversampling uses SMOTE from imbalanced-learn [21]. Indonesian stopword removal uses the Sastrawi library, and figures use Matplotlib and Seaborn. A fixed random seed of 42 was used throughout for reproducibility.

2.5 Research Procedures or Stages

In the first stage, the raw dataset initially comprised 457 reviews across a seven-class distribution. To resolve data hygiene issues and prevent potential data leakage, duplicate reviews were systematically removed, yielding a cleaned dataset of 446 unique instances. This raw distribution exhibited severe class imbalance, where the three smallest minority classes contained fewer samples than the default k-nearest neighbors parameter required by the SMOTE algorithm. To mitigate this computational constraint, the classes fear, surprise, and disgust were consolidated into a single new category designated as OTHER. This adjustment resulted in a refined five-class dataset (N=446).

During the second stage, the raw text underwent a bifurcated preprocessing pipeline to accommodate the specific input requirements of the different algorithms. The first version, designed for IndoBERT, applied light cleaning procedures that included lowercasing, the removal of URLs, mentions, and emojis, alongside slang normalization utilizing a custom dictionary incorporating local Sundanese terms. The second version, tailored for TF-IDF and exploratory data analysis, built upon the light cleaning by further removing punctuation and digits, and applying Indonesian stopword removal via the Sastrawi library. Crucially, negation words such as *tidak*, *kurang*, *jangan*, *belum*, and *bukan* were intentionally retained during the stopword removal phase to preserve the original emotional polarity of the reviews.

As illustrated in the subsequent stages, the pipeline then utilized a frozen IndoBERT extractor to generate a 446 x 768 feature matrix from the [CLS] tokens. Finally, the modeling and evaluation phase implemented a 5-Fold Stratified Cross-Validation strategy to rigorously evaluate the models. For the M3 variant, Principal Component Analysis (PCA) was incorporated prior to SMOTE to safely reduce the high-dimensional embeddings (n=50) and stabilize the synthetic oversampling process.

Following the text preprocessing stage, feature extraction was conducted using IndoBERT, which was deployed strictly as a frozen feature extractor without any fine-tuning. Each review was tokenized to a maximum length of 256 tokens, and the final hidden state of the [CLS] token was extracted to represent the entire text sequence. As detailed in the architectural flow in Fig. 2, this process generated a 768-dimensional embedding vector for each document, resulting in a finalized feature matrix of 446 by 768 that was cached for computational efficiency.

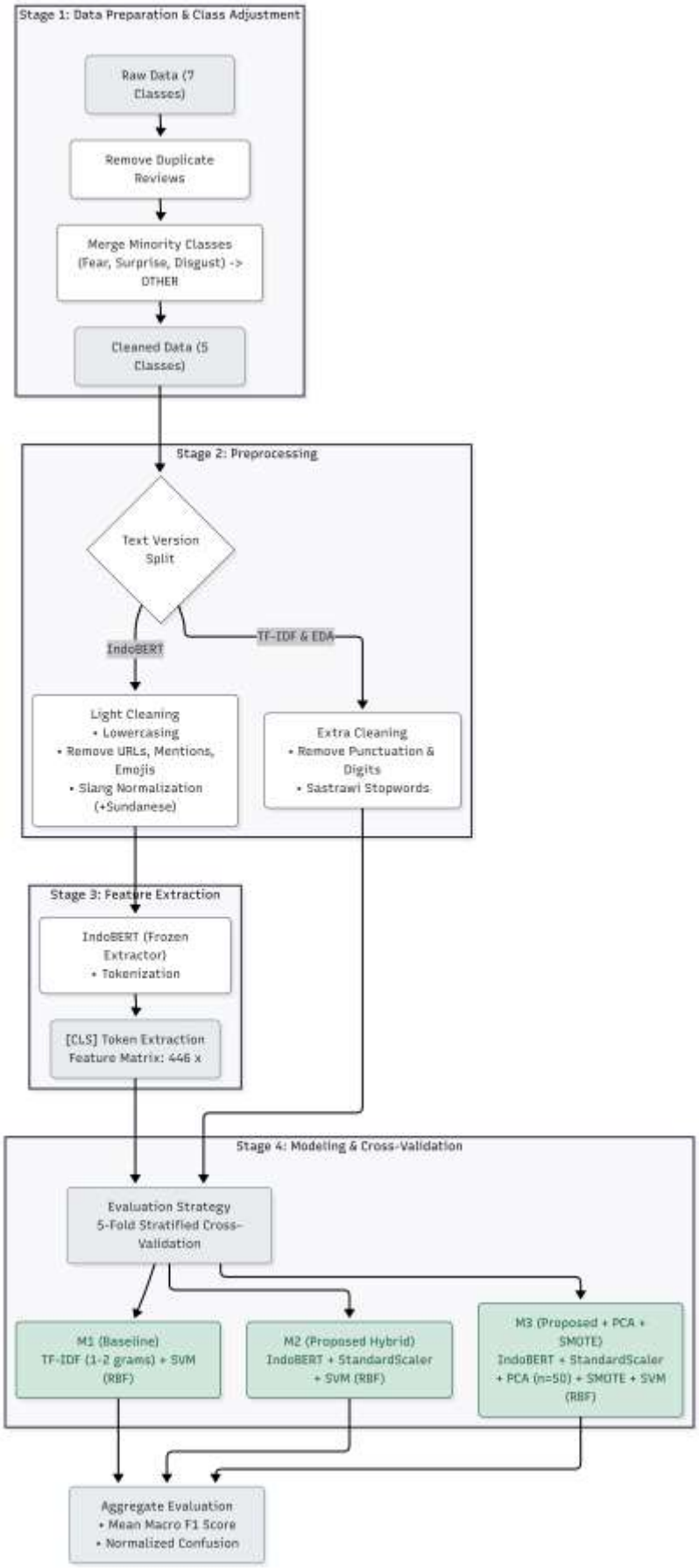


Fig. 1. Research pipeline from data collection to analysis.

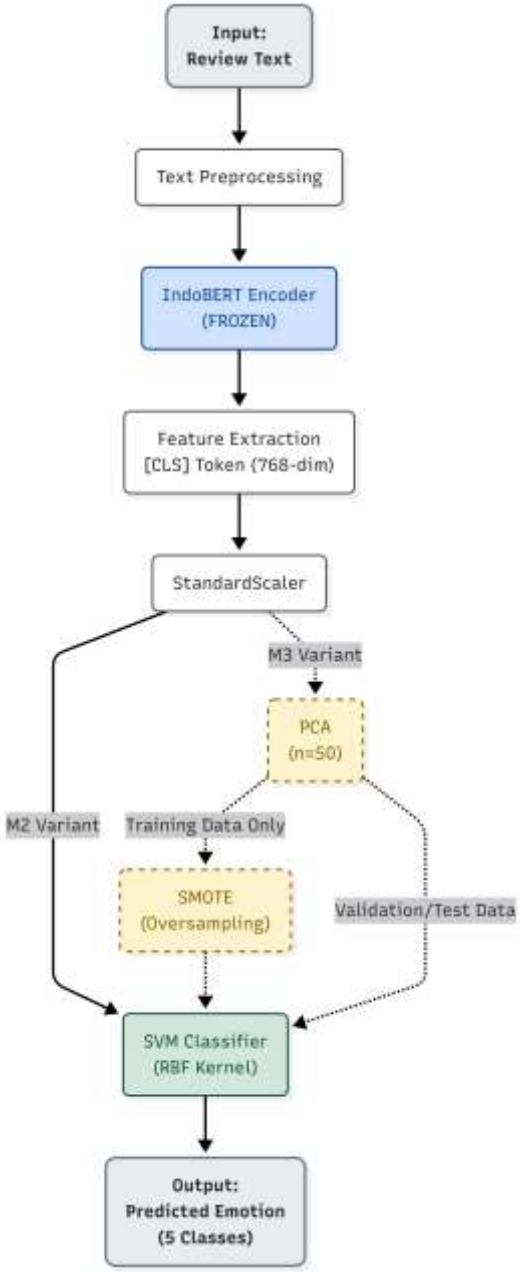


Fig. 2. Hybrid IndoBERT and SVM architecture. SMOTE applies to the M3 variant only and is fitted on training data inside the pipeline.

To evaluate classification performance, three distinct models were trained. Ensuring an equitable and stable performance comparison across the heavily imbalanced data, the evaluation utilized a 5-Fold Stratified Cross-Validation framework rather than a single static data split. The baseline model, designated as M1, utilized TF-IDF vectorization—incorporating unigrams and bigrams, a minimum document frequency of two, and sublinear term frequency—coupled with a Support Vector Machine (SVM) employing a Radial Basis Function (RBF) kernel set to C = 10 and gamma = scale. The proposed hybrid model, M2, standardized the extracted IndoBERT [CLS] features via a StandardScaler before passing them directly to the identical SVM classifier.

To address the remaining class imbalance, the M3 variant integrated a specialized imbalanced-learn pipeline. As illustrated in the branching logic of Fig. 2, the standardized IndoBERT features in the M3 route first underwent Principal Component Analysis (PCA) to safely reduce the high-dimensional embeddings to 50 components. Following this dimensionality reduction, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Crucially, by enclosing these steps within the cross-validation pipeline, the architecture guarantees strict validation integrity; PCA parameters are derived solely from the training folds

and applied to both training and validation sets, while SMOTE is executed exclusively on the training folds. This rigorous isolation explicitly prevents any synthetic data leakage into the validation sets, yielding highly reliable out-of-fold predictions for the final SVM classifier.

2.6 Data Analysis Techniques

To facilitate a rigorous and unbiased comparative analysis across the three evaluated models, the experimental framework employed a 5-Fold Stratified Cross-Validation strategy on the curated dataset of 446 reviews, thereby mitigating the statistical variance inherent in conventional single-split validation. The models' predictive capabilities were assessed utilizing several metrics aggregated across all folds: mean accuracy, per-class precision, recall, and F1-score, alongside mean macro F1, weighted F1, and normalized confusion matrices derived from out-of-fold predictions. The mean macro F1-score was designated as the primary evaluation metric; by computing the unweighted mean of class-specific scores, this metric systematically prevents the artificial inflation of model efficacy by majority classes, which is an indispensable requirement when evaluating heavily imbalanced datasets. Furthermore, the inter-rater reliability of the manual annotation process was quantified utilizing Cohen's Kappa coefficient, with the resulting agreement thresholds interpreted according to the Landis and Koch benchmark [22]. Finally, the spatial emotion profiling was conducted via a cross-tabulation of the tourist destinations against the predicted emotion labels. However, given the highly constrained sample sizes corresponding to specific locales, this destination-level analysis is strictly delimited to an exploratory scope rather than serving as a foundation for definitive managerial interventions.

3. Results and Discussion

3.1 Presentation of Research Results

Annotation reliability. Following the removal of duplicate entries to ensure strict data hygiene, the manual human annotations and the independent LLM validations agreed on 379 out of the 446 unique reviews. The observed proportional agreement is po =0.8498, while the expected agreement by chance is pe = 0.4051. Cohen's Kappa is therefore (0.8498 minus 0.4051) divided by (1 minus 0.4051), which equals 0.7475. According to the Landis and Koch scale, this coefficient signifies substantial agreement, thereby validating the robustness of the ground truth labels established prior to model training. The comprehensive cross-tabulation of the agreement matrix between the human researcher (rows) and the LLM validator (columns) is detailed in Table 1.

Table 1. Agreement matrix between researcher labels (rows) and LLM labels (columns), N = 446. Diagonal cells are agreements.

Researcher	LLM							
	ANGER	***DISGUST***	***FEAR***	***JOY***	***NEUTRAL***	***SADNESS***	***SURPRISE***	***Total***
ANGER	10	0	0	1	1	2	0	14
DISGUST	0	6	0	0	1	1	0	8
FEAR	0	0	3	1	0	1	0	5
JOY	1	0	0	231	18	5	0	255
NEUTRAL	1	0	0	18	85	2	0	106
SADNESS	0	0	0	7	6	38	0	51
SURPRISE	0	0	0	1	0	0	6	7
Total	12	6	3	259	111	49	6	446

Data distribution. Table 2 delineates the class frequencies of the dataset before and after the consolidation of the three smallest minority classes. The curated dataset of 446 unique reviews initially exhibited a severe class imbalance distributed across seven emotional categories: JOY (255), NEUTRAL (106), SADNESS (51), ANGER (14), DISGUST (8), SURPRISE (7), and FEAR (5). The extreme scarcity of instances within the latter three categories presented a critical methodological constraint, as oversampling algorithms such as SMOTE require a sufficient basal support size to reliably compute the *k*-nearest neighbors for synthetic data generation. To circumvent this computational bottleneck, instances labeled as DISGUST, SURPRISE, and FEAR were aggregated into a unified 'OTHER' category, accumulating a combined total of 20 samples. This strategic merging yielded a refined five-class distribution. Although the inherent imbalance of the dataset naturally reflects the highly skewed polarity typical of tourism reviews, this structural adjustment establishes a

mathematically viable foundation for the subsequent oversampling procedures and ensures stable model optimization.

Table 2. Class distribution before and after merging the three smallest classes into OTHER

Class (7)	Count	Class (5)	Count
JOY	255	JOY	255
NEUTRAL	106	NEUTRAL	106
SADNESS	51	SADNESS	51
ANGER	14	ANGER	14
DISGUST	8	OTHER	20
SURPRISE	7		
FEAR	5		

To visually conceptualize the outcome of this exploratory analysis prior to feature extraction and model training, the structural shift from the initial seven distinct categories to the finalized five modeling classes is illustrated in Fig. 3. This graphical representation clearly contrasts the heavy dominance of the majority class against the newly aggregated 'OTHER' category, visually reinforcing the necessity of the aforementioned class-merging strategy to mitigate computational risks during the subsequent oversampling phase.

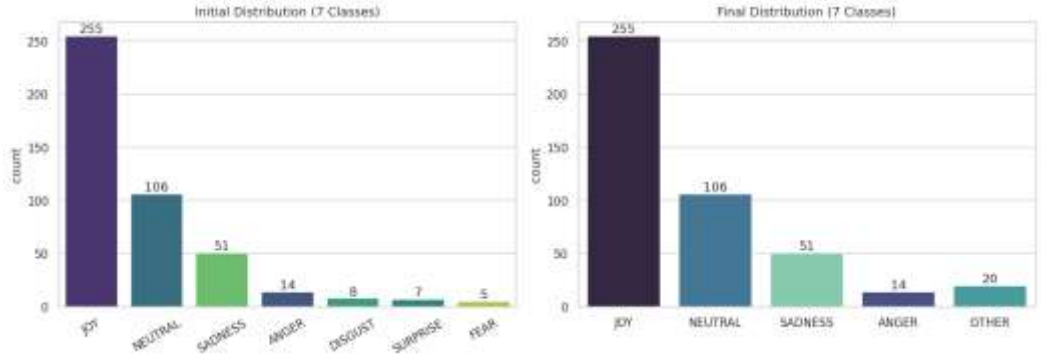


Fig. 3. Emotion class distribution before (left) and after (right) the minority classes merging.

Three further views describe the corpus. Fig. 4 shows the distribution of review length in words, which is right skewed, together with length per class, where the boxes overlap so length on its own does not separate the emotions. Fig. 5 shows the most frequent words in each class after stopword removal, which gives a qualitative sense of the vocabulary tied to each emotion. Fig. 6 shows how the reviews split across destinations, which is uneven and concentrated at Batu Alam.

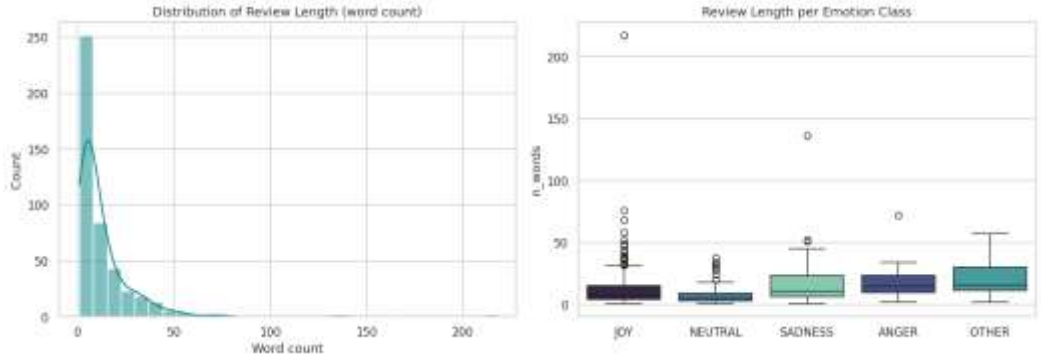


Fig. 4. Review length in words across the corpus on the left and per emotion class on the right.

Fig. 5. Word clouds for each of the five emotion classes after stopword removal.

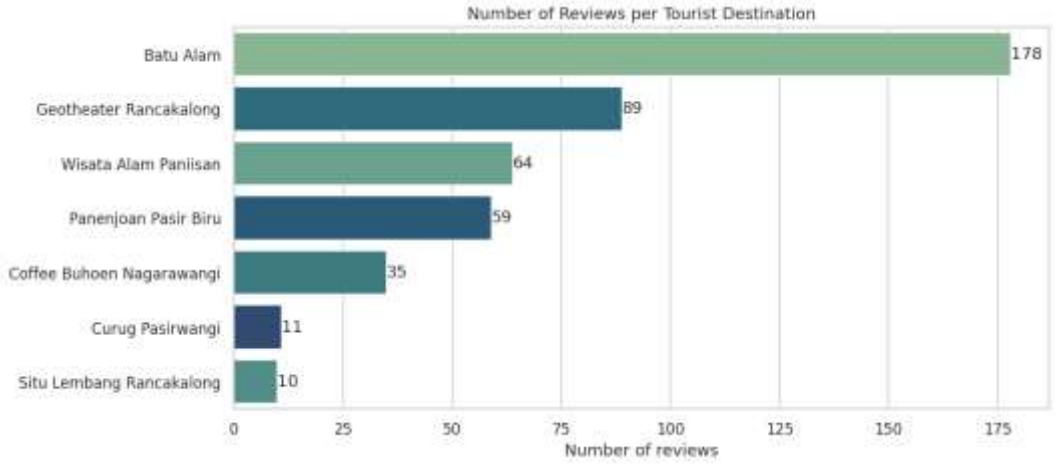


Fig. 6. Number of reviews collected per destination.

Table 3 presents the aggregate performance metrics of the three evaluated architectures across the 5-Fold Stratified Cross-Validation framework. A comparative analysis of these results highlights the distinct impact of both the feature extraction techniques and the class-balancing interventions. The baseline model (M1), which relies on lexical TF-IDF features, achieved an accuracy of 0.6816. However, its low Macro F1 score of 0.3920 exposes a severe susceptibility to the dataset's inherent imbalance, indicating that the model heavily favored the majority class while struggling to accurately classify minority instances.

Transitioning to the contextual embeddings provided by the frozen IndoBERT extractor, the M2 variant demonstrated a marked improvement. M2 achieved the highest overall accuracy among the configurations at 0.7131, alongside an increased Macro F1 score of 0.4133. This indicates that the dense, 768-dimensional semantic representations successfully captured deeper linguistic nuances compared to traditional sparse vectorization, thereby enhancing general predictive capabilities.

Nevertheless, overall accuracy is a fundamentally deceptive metric in highly skewed datasets. The proposed M3 architecture was explicitly engineered to optimize predictive parity across all emotional categories by integrating PCA and SMOTE. While M3 experienced a slight reduction in aggregate accuracy (0.6839) compared to M2—a mathematically expected trade-off when forcing a classifier to pay equal attention to rare classes—it successfully achieved the highest Macro F1 score at 0.4321. Because Macro F1 calculates the unweighted mean of class-specific metrics, it serves as the most unbiased indicator of model robustness. The superiority of M3 in this primary metric empirically validates that the combination of dimensionality reduction and synthetic oversampling effectively strengthened the Support Vector Machine's decision boundaries, making it the most balanced and reliable configuration for classifying this imbalanced tourism review dataset.

Table 3. Performance comparison of the three models using 5-Fold Cross-Validation (N = 446).

Model	Accuracy	Macro precision	Macro recall	Macro F1	Weighted F1
M1: SVM + TF-IDF	0.6816	0.5352	0.3683	0.3920	0.6349
M2: IndoBERT + SVM	0.7131	0.4448	0.4140	0.4133	0.6831
M3: IndoBERT + SVM + SMOTE	0.6839	0.5040	0.4195	0.4321	0.6649

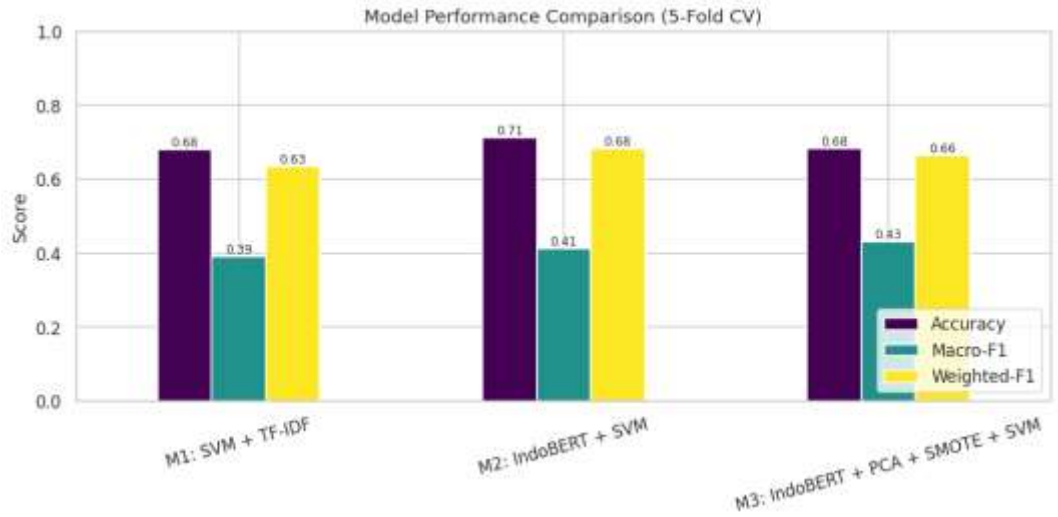


Fig. 7. Performance comparison of the three evaluated models across Accuracy, Macro-F1, and Weighted-F1 metrics.

This intricate performance dynamic is visually corroborated by Fig. 7, which illustrates the comparative scores of the three models across the evaluation criteria. The visual representation clearly demonstrates that the model rankings are not uniformly consistent across all metrics. As depicted in the bar chart, the M2 architecture achieves the highest values in both Accuracy (0.71) and Weighted-F1 (0.68)—metrics that disproportionately reward correct predictions within the dominant majority class. However, M3 deliberately sacrifices a fraction of these majority-biased metrics to attain the highest Macro-F1 score (0.43), thereby proving its enhanced sensitivity to minority classes. Meanwhile, the baseline M1 configuration remains demonstrably inferior, occupying the lowest tier across all evaluated metrics (Accuracy: 0.68; Macro-F1: 0.39; Weighted-F1: 0.63). Ultimately, Fig. 7 reinforces the methodological imperative of relying on Macro-F1 to ascertain true model efficacy when handling severely imbalanced datasets.

Per-class F1 results elucidate the specific areas where architectural adjustments affect predictive performance. Table 4 details the out-of-fold per-class F1-scores for all three evaluated configurations. Under the M1 baseline, the model completely fails to identify the minority ANGER class (F1 = 0.00), although it achieves a modest score of 0.26 on the aggregated OTHER category. Transitioning to the M2 IndoBERT model yields notable improvements across the more populated classes, with JOY increasing to 0.82, NEUTRAL to 0.63, and SADNESS to 0.53. Furthermore, M2 initiates a slight recovery for the ANGER class (F1 = 0.12). However, the dense semantic representations of M2 struggle significantly with the OTHER class, dropping its performance precipitously to 0.00.

The application of PCA and SMOTE in the M3 variant produces mixed but theoretically consistent outcomes. While SADNESS reaches its peak F1-score (0.55) and the OTHER class partially recovers to 0.14, the metric for ANGER remains persistently low at 0.11, and NEUTRAL exhibits a regression to 0.56. It is crucial to contextualize these fluctuations against the absolute support sizes. With ANGER and OTHER possessing total occurrences of only 14 and 20 instances respectively across the entire dataset, their evaluation metrics remain statistically fragile. In such extremely constrained minority classes, the misclassification of merely one or two instances translates to drastic percentage shifts, limiting the magnitude of improvement SMOTE can reliably yield.

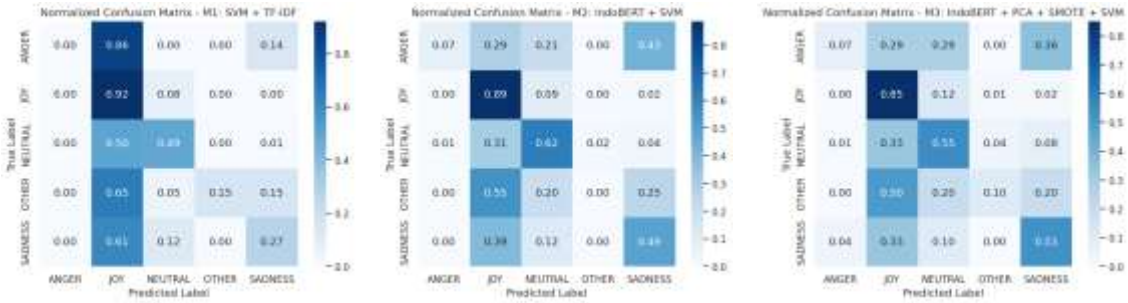
Finally, the persistent instability and generally poor performance on the OTHER class across both advanced models confirm that consolidating distinct, unrelated emotions (fear, surprise, disgust) into a single category was ultimately suboptimal. The inability of the Support Vector Machine to establish a consistent decision boundary over these merged instances suggests that this catch-all class inherently lacks a coherent and unified semantic signature.

Table 4. Agregate out-of-fold per-class F1-scores for M1, M2, and M3 evaluated via 5-fold-cross-validation.

Class	Support	M1 F1	M2 F1	M3 F1
ANGER	14	0.00	0.12	0.11
JOY	255	0.78	0.82	0.81
NEUTRAL	106	0.56	0.63	0.56
OTHER	20	0.26	0.00	0.14
SADNESS	51	0.39	0.53	0.55

The normalized confusion matrices presented in Fig. 8 visually corroborate these statistical observations by mapping the exact migration of predictions across the models. As depicted in Fig. 8a, the baseline M1 model exhibits extreme majority bias, misclassifying 86% of the true ANGER instances as JOY and completely failing to identify the ANGER class correctly (0.00). Transitioning to the M2 IndoBERT architecture in Fig. 8b, the model begins to faintly detect ANGER (0.07) but simultaneously loses all capacity to identify the OTHER class (0.00), misallocating the majority of its instances (55%) into the dominant JOY category.

The application of PCA and SMOTE in Fig. 8c illustrates a nuanced architectural trade-off. While SMOTE successfully strengthens the true positive rates for SADNESS (0.53) and partially recovers the OTHER class (0.10), it fails to substantially improve the detection of ANGER, which remains stagnant at 0.07. To accommodate these minority improvements, the M3 model visibly sacrifices a portion of its true positive rate for the dominant JOY class, which drops from 0.89 in M2 to 0.85 in M3. Ultimately, the persistent misclassification of the OTHER instances across all three matrices—primarily bleeding into JOY and NEUTRAL—serves as explicit visual proof that synthetically merging heterogeneous emotions prevents the algorithm from establishing functional, discernible decision boundaries.



(a) (b) (c)

Fig. 8. Normalized confusion matrix for M1(a), M2(b), and M3(c)

To demonstrate the practical utility of the optimal M3 classifier, an exploratory spatial emotion profiling was conducted across the seven tourist locations in the Rancakalong region. Table 5 details the absolute cross-tabulation of destinations against the predicted emotion labels, while Fig. 9 visualizes this distribution as proportional stacked bar charts.

Table 5. Number of reviews per destination and emotion class.

Destination	JOY	NEUTRAL	SADNESS	ANGER	OTHER	Total
Batu Alam	124	31	10	8	5	178
Geotheater Rancakalong	38	27	18	1	5	89
Wisata Alam Paniisan	36	20	6	1	3	66
Panenjoan Pasir Biru	29	9	11	4	6	59
Coffee Buhoen Nagarawangi	20	13	2	0	0	35
Curug Pasirwangi	5	2	3	0	1	11
Situ Lembang Rancakalong	5	4	1	0	0	10

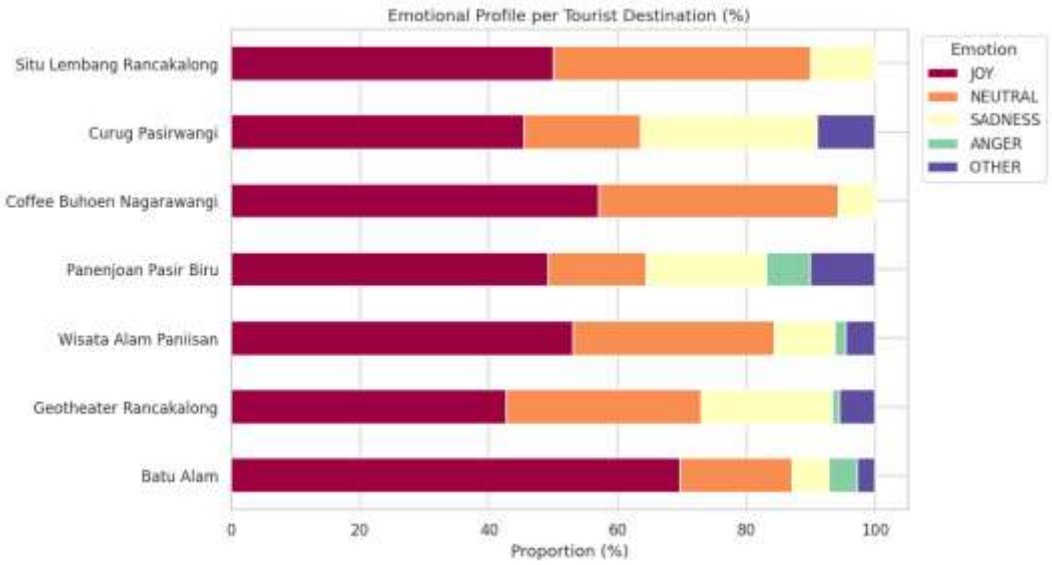


Fig. 9. Proportional emotion profile for each tourist destination based on the predicted review labels.

Unsurprisingly, 'JOY' operates as the dominant baseline emotion across all locations, validating the general satisfaction typical of tourism experiences. Batu Alam, possessing the highest data support (N=178), exhibits a robust positive profile, heavily skewed towards JOY. However, when normalizing the data into proportions (Fig. 9), distinct negative emotional signatures emerge in specific locales. Geotheater Rancakalong and Panenjoan Pasir Biru display visibly thicker bands of 'SADNESS' relative to their total reviews. Similarly, the 'ANGER' proportion is most pronounced at Panenjoan Pasir Biru.

It is imperative to note that for destinations such as Coffee Buhoen Nagarawangi, Curug Pasirwangi, and Situ Lembang Rancakalong, the review volume is extremely sparse (N<35). Consequently, the proportional bars for these specific locations are statistically sensitive and should not be over-interpreted. Nevertheless, for the locations with adequate sample sizes, these localized emotional profiles offer a preliminary diagnostic tool, directing local agro-tourism management toward the specific sites that may require infrastructural or service-level interventions.

3.2 Analysis of Findings

The experimental results directly address the first research question regarding the efficacy of contextual feature representation. Replacing the sparse TF-IDF vectors with dense, frozen IndoBERT embeddings (M2) increased the overall accuracy from 0.6816 to 0.7131, alongside a positive shift in the Macro F1 score from 0.3920 to 0.4133. The fundamental driver of this improvement is the capture of semantic context. TF-IDF treats reviews as disconnected bags of words, which yields weak and sparse signals for short texts containing informal or rare vocabulary. Conversely, the 768-dimensional IndoBERT [CLS] token encodes word order and contextual nuance, enabling the SVM classifier to better distinguish underlying emotions that share surface-level words. This finding corroborates prior literature demonstrating the superiority of contextual pretraining over traditional frequency-based features for Indonesian short-text classification [6], [7].

The second research question investigates the capability of oversampling techniques under extreme class imbalance. Integrating PCA dimensionality reduction alongside SMOTE (M3) successfully elevated the Macro F1 score to its peak at 0.4321. Notably, this gain in Macro F1 was accompanied by a deliberate reduction in aggregate accuracy, which dropped to 0.6839. Rather than a deficiency, this pattern exemplifies the mathematically expected trade-off inherent in balancing algorithms: the model explicitly sacrifices a degree of precision within the heavily dominant 'JOY' class to establish more equitable and sensitive decision boundaries for the minority classes [12], [21]. The fact that the primary improvement manifests in Macro F1 empirically validates that the M3 architecture is the most robust and unbiased configuration for highly skewed datasets, as accuracy metrics are inherently misleading when a single class dominates the distribution.

The per-class evaluation further delineates the operational cost of this imbalance. The baseline M1 model completely failed to identify the ANGER class (F1 = 0.00), defaulting its predictions to the frequent categories. While M2 and M3 initiated partial recoveries for ANGER (F1 = 0.12 and 0.11, respectively), these gains remained constrained by the extremely small absolute support size ($N=14$), where a single misclassification causes severe statistical volatility. Furthermore, the erratic performance on the 'OTHER' class—which scored 0.26 in M1, plummeted to 0.00 in M2, and partially recovered to 0.14 in M3—highlights the fundamental limitation of merging heterogeneous minority classes. Because 'OTHER' is a synthetic amalgamation of fear, surprise, and disgust, it lacks a cohesive linguistic signature. Even with SMOTE

densifying the minority feature space, the SVM struggled to map a consistent, functional boundary for this catch-all category.

Finally, the spatial destination analysis answers the third research question while fulfilling the primary output of the associated community service project in Rancakalong, which strategically prioritizes analytical model development over the deployment of a full-scale information system. Batu Alam constitutes the largest share of the dataset (178 reviews) and is overwhelmingly characterized by joy (124). However, the optimal M3 model successfully isolated negative emotional pockets: Geotheater Rancakalong and Panenjoan Pasir Biru carry a disproportionately larger share of sadness (18 of 89, and 11 of 59, respectively), alongside notable concentrations of anger (4 at Panenjoan Pasir Biru, 8 at Batu Alam). While smaller sites like Curug Pasirwangi (11) and Situ Lembang (10) possess insufficient data to draw definitive operational conclusions, the spatial emotion profiling generated by the model effectively highlights specific tourist locales that warrant targeted managerial evaluation.

3.3 Implications of the Results

From a practical perspective, the spatial emotion profiles furnish local stakeholders with a diagnostic metric to identify specific clusters of visitor dissatisfaction, thereby guiding targeted infrastructural and service interventions. By transforming unstructured, publicly available textual data into actionable insights, this research serves as a foundational analytical model for agro-tourism development in Rancakalong. This localized computational approach directly supports the sustainable economic growth and community resilience targets outlined in SDG 8 and SDG 11, effectively providing data-driven governance without necessitating the immediate deployment of a complex, full-scale information system.

Methodologically, this study establishes empirical evidence that integrating a frozen IndoBERT feature extractor with a Support Vector Machine offers a highly robust and computationally efficient framework for classifying limited Indonesian text data, particularly when full large language model fine-tuning is resource-prohibitive. Crucially, the research highlights the absolute necessity of rigorous validation architectures when handling extreme class imbalance. By confining dimensionality reduction (PCA) and synthetic oversampling (SMOTE) strictly within the training folds of the cross-validation pipeline, this study successfully demonstrates how to mitigate class skewness while definitively preventing the synthetic data leakage that frequently invalidates naive oversampling approaches.

Furthermore, the rigorously curated dataset constitutes a significant academic contribution to the field of natural language processing in its own right. It delivers an original, duplicate-free, and LLM-validated emotion corpus comprising 446 unique Indonesian tourism reviews. The dataset's ground truth is mathematically bolstered by a substantial inter-rater reliability consensus, evidenced by a Cohen's Kappa coefficient of 0.7475, thereby providing a highly reliable benchmark for future emotion classification research in low-resource linguistic contexts.

3.4 Limitations of the Study

While the revised methodology systematically addresses algorithmic vulnerabilities through rigorous cross-validation and dimensionality reduction, this study acknowledges several persistent limitations. First, the structural consolidation of distinct minority emotions (fear, surprise, and disgust) into a unified 'OTHER' category proved to be a suboptimal semantic compromise. Although this aggregation provided the necessary mathematical basal support for the k-nearest neighbor computations during oversampling, the resulting class remained excessively heterogeneous. This lack of a cohesive linguistic signature prevented the classifier from establishing a reliable decision boundary, indicating that merging inherently different emotions is a flawed strategy for fine-grained sentiment analysis.

Second, despite the implementation of robust 5-Fold Stratified Cross-Validation and the application of Principal Component Analysis (PCA) to stabilize the high-dimensional IndoBERT embeddings prior to SMOTE, the absolute scarcity of minority samples persists as a statistical constraint. Within the curated dataset of 446 unique reviews, categories such as ANGER possess an aggregate support of only 14 instances. At this minimal scale, per-class evaluation metrics are inherently fragile; isolated misclassifications disproportionately skew precision and recall scores, thereby constraining the statistical confidence of the model's performance on exceedingly rare emotions.

Finally, the compiled corpus is geographically restricted to specific agro-tourism destinations in Rancakalong and is strongly influenced by localized linguistic patterns, including informal Sundanese vernacular. Consequently, the established baseline models and spatial emotion profiles may lack direct generalizability to broader Indonesian tourism contexts. Future research should prioritize corpus expansion across diverse geographic regions and explore advanced algorithmic handling of extreme minority classes without relying on synthetic aggregation.

4. Conclusion

This study successfully developed a robust multi-class emotion classification framework for Indonesian agro-tourism reviews by integrating a frozen IndoBERT feature extractor with a Support Vector Machine (SVM). Addressing severe class imbalance within a rigorously curated corpus of 446 unique reviews, the research implemented a 5-Fold Stratified Cross-Validation pipeline to ensure unbiased evaluation. The empirical results demonstrate that replacing traditional TF-IDF sparse vectors with dense IndoBERT contextual embeddings increased overall predictive accuracy from 0.6816 to 0.7131, alongside an improvement in the Macro F1 score from 0.3920 to 0.4133. To further counteract majority class bias without inducing data leakage, the optimal M3 architecture integrated Principal Component Analysis (PCA) and the SMOTE algorithm exclusively within the training folds. This sophisticated configuration achieved the highest aggregate Macro F1 score of 0.4321, empirically validating that dimensionality reduction combined with synthetic oversampling establishes the most equitable decision boundaries for minority classes.

Despite these algorithmic triumphs, the persistent statistical fragility and erratic predictive performance of the consolidated 'OTHER' category underscore a critical methodological lesson: merging inherently distinct, rare emotions (fear, surprise, disgust) into a synthetic catch-all class is fundamentally suboptimal, as it deprives the classifier of a coherent semantic signature. Practically, the finalized model generates spatial emotion profiles that provide site managers in Rancakalong with a viable, data-driven diagnostic tool, fulfilling the core analytical objectives of the targeted community service initiative. Moving forward, future research must prioritize substantial localized corpus expansion to naturally satisfy algorithmic support requirements, thereby eliminating the reliance on synthetic class aggregation and facilitating the full-scale fine-tuning of large language models.

Acknowledgment

The authors thank the reviewers and the contributors who collected and annotated the dataset.

Declarations

Author contribution. All authors contributed to the study design, analysis, and writing, and approved the final manuscript
Funding statement. The research received no specific grant.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The dataset (Dataset_Wisata_RancakalongSumedang-humanv3.csv), the annotation prompt, the Cohen's Kappa computation, and the analysis notebook (v1code-run Klasifikasi_Emosi_IndoBERT_SVM.ipynb) are available in the project repository. The language model indobenchmark/indobert-base-p1 is publicly available through the Hugging Face model hub.

References

[1] Y. Mao, Q. Liu, and Y. Zhang, “Sentiment analysis methods, applications, and challenges: A systematic literature review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, p. 102048, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.

[2] P. Ekman, “An argument for basic emotions,” *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[3] R. Plutchik, “A general psychoevolutionary theory of emotion,” in *Emotion: Theory, Research, and Experience*, vol. 1, R. Plutchik and H. Kellerman, Eds., New York: Academic Press, 1980, pp. 3–33.

[4] Y. Kurniawati, R. B. Hamid, D. I. Sensuse, S. Lusa, P. A. W. Putro, and S. Indriasari, “Analysis of Public Sentiment Indonesia’s Personal Data Protection Law: A Comparison of SVM and IndoBERT on X Platform,” *J. Tek. Inform. (JUTIF)*, vol. 7, no. 2, pp. 1007–1027, Apr. 2026, doi: 10.52436/1.jutif.2026.7.2.5415.

[5] A. Vaswani *et al.*, “Attention Is All You Need,” 2017, *arXiv*. doi: 10.48550/ARXIV.1706.03762.

[6] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, 2020, pp. 843–857.

[7] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020, pp. 757–770.

[8] F. Koto, J. H. Lau, and T. Baldwin, “IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization,” in *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 10660–10668.

[9] A. R. Setyawan, L. H. Suadaa, and B. Yuniarto, "Aspect-Based Sentiment Analysis using Adaptive Aspect on Tourist Reviews in Jakarta," *SISTEMASI*, vol. 13, no. 6, p. 2456, Nov. 2024, doi: 10.32520/stmsi.v13i6.4585.

[10] E. Junianto, M. Puspitasari, S. I. Zakaria, T. Arifin, and I. W. P. Agung, "Emotion Detection in Indonesian Text Using the Logistic Regression Method," *media j. inform.*, vol. 17, no. 2, pp. 305–316, Dec. 2025, doi: 10.35194/mji.v17i2.5927.

[11] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *JCCE*, vol. 2, no. 4, pp. 352–356, Apr. 2023, doi: 10.47852/bonviewJCCE3202838.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[13] A. A. Lestari, Ahmad Faqih, and Gifthera Dwilestari, "Improving Sentiment Analysis Performance of Tokopedia Reviews Using Principal Component Analysis and Naïve Bayes Algorithm," *j. of artif. intell. and eng. appl.*, vol. 4, no. 2, pp. 758–763, Feb. 2025, doi: 10.59934/jaiea.v4i2.743.

[14] S. Syah Putra and D. Riminarsih, "Evaluating Machine Learning Models Across Feature Extraction and Data Balancing Scenarios for Coretax Sentiment Analysis," *media j. inform.*, vol. 17, no. 2, pp. 379–396, Dec. 2025, doi: 10.35194/mji.v17i2.5968.

[15] F. Nurpandi, F. S. Sulaeman, and A. Hermawan, "Analisis Sentimen Terhadap Kinerja Kepolisian Indonesia Menggunakan Metode Multinomial Naive Bayes, Long Short-Term Memory, dan Lexicon-Based," *MJI*, vol. 16, no. 1, p. 1, Jun. 2024, doi: 10.35194/mji.v16i1.4165.

[16] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877–1901.

[17] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, pp. 27730–27744.

[18] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, p. e2305016120, 2023.

[19] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2020, pp. 38–45.

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[22] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.