VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7729

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

LEVERAGING CONTINUAL FINE-TUNING FOR EMOTION CLASSIFICATION IN PRODUCT REVIEWS ON MSME SUSTAINABILITY SUPPORT

Galih Setiawan Nurohim[1*]; Heribertus Ary Setyadi[1]; Pudji Widodo[2]; Yusuf Sutanto[3]

Information System[1]
Computer Technology[2]
Universitas Bina Sarana Informatika, Indonesia[1,2]
https://bsi.ac.id[1,2]
galih.glt@bsi.ac.id*, heribertus.hbs@bsi.ac.id, pudji.pwd@bsi.ac.id

Informatics[3]
Universitas Dharma AUB, Indonesia[3]
https://undha.ac.id[3]
yusuf.sutanto@stie-aub.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Automatic analysis of consumer product reviews is essential for understanding granular customer perceptions beyond basic sentiment. While transformer-based models are prevalent in Indonesian sentiment analysis, their adaptation for multi-emotion classification shifting from broad polarities to specific affective states remains underexplored. This study addresses this gap by proposing a Continual Fine-Tuning (CFT) approach to adapt a pre-trained IndoBERTweet model from three sentiment categories into five distinct emotion classes: Happiness, Sadness, Fear, Love, and Anger. The novelty lies in the strategic repurposing of sentiment-oriented weights to capture nuanced emotional representations in Indonesian e-commerce discourse. Experimental results on the PRDECT-ID dataset demonstrate that the proposed CFT model achieves an accuracy of 0.8157 and a weighted F1-score of 0.8118, significantly outperforming traditional neural networks and multilingual baselines. The CFT model demonstrates a 2.13% improvement in accuracy compared to the base IndoBERTweet without continual tuning and a substantial 59.54% lead over the multilingual BERT (mBERT) baseline. Despite limitations concerning the dataset scale (5,400 samples) and inherent subjectivity in emotion labeling, this research provides a robust conceptual framework for model adaptation in the Indonesian NLP ecosystem. These findings suggest that CFT is an efficient strategy for enhancing the emotional intelligence of transformer models, especially in domain-specific tasks where high-quality labeled data is constrained.*

Keywords: *Continual Fine-Tuning, Emotion Classification, Indonesian NLP, IndoBERTweet, Transformer Adaptation.*

Intisari— *Analisis otomatis terhadap ulasan produk konsumen sangat penting untuk memahami persepsi pelanggan yang lebih mendalam melampaui polaritas sentimen dasar. Meskipun model berbasis transformer telah banyak digunakan dalam analisis sentimen bahasa Indonesia, adaptasinya untuk klasifikasi multi-emosi—beralih dari polaritas luas ke status afektif yang spesifik—masih jarang dieksplorasi. Penelitian ini menjawab celah tersebut dengan mengusulkan pendekatan Continual Fine-Tuning (CFT) untuk mengadaptasi model IndoBERTweet yang telah dilatih sebelumnya dari tiga kategori sentimen menjadi lima kelas emosi yang berbeda: Senang (Happiness), Sedih (Sadness), Takut (Fear), Cinta (Love), dan Marah (Anger). Kebaruan penelitian ini terletak pada pemanfaatan kembali secara strategis (strategic repurposing) bobot model yang berorientasi sentimen untuk menangkap representasi emosional yang bernuansa dalam diskursus e-*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7729

commerce di Indonesia. Hasil eksperimen pada dataset PRDECT-ID menunjukkan bahwa model CFT yang diusulkan mencapai akurasi sebesar 0,8157 dan skor F1 tertimbang sebesar 0,8118, mengungguli jaringan saraf tradisional dan baseline multibahasa secara signifikan. Meskipun terdapat keterbatasan terkait skala dataset (5.400 sampel) dan subjektivitas inheren dalam pelabelan emosi, penelitian ini menyediakan kerangka kerja konseptual yang kuat untuk adaptasi model dalam ekosistem NLP bahasa Indonesia. Temuan ini menunjukkan bahwa CFT merupakan strategi yang efisien untuk meningkatkan kecerdasan emosional model transformer, terutama dalam tugas-tugas spesifik domain dengan keterbatasan data berlabel berkualitas tinggi.

***Kata Kunci**: Continual Fine-Tuning, IndoBERTweet, Klasifikasi Emosi, NLP Bahasa Indonesia, Adaptasi Transformer.*

INTRODUCTION

In the contemporary period, many individuals are establishing Micro, Small, and Medium-sized Enterprises (MSMEs) as a means of earning an income and mitigating poverty and unemployment, which has led to the proliferation of numerous new MSMEs [1]. A large number of MSMEs are utilizing the internet, particularly through e-commerce platforms and social media, to market their products and expand their market reach [2]. After completing a purchase, consumers are often encouraged to submit product reviews expressing their opinions about the purchased items [3]. These reviews can serve as indicators of customer satisfaction and may influence the purchasing decisions of other potential buyers [4]. Therefore, it is important for sellers to analyze customer reviews and understand the emotions expressed in them in order to improve their products and services [5]. According to data provided by Statista Market Insights, Indonesia's e-commerce user base amounted to 178.94 million in 2022, with a projection to reach 196.47 million users by the close of 2023 [6].

Recent advances in artificial intelligence have significantly transformed the field of Natural Language Processing (NLP), particularly with the emergence of Large Language Models (LLMs) capable of understanding complex linguistic patterns. These models have demonstrated strong performance across various NLP tasks, including text classification, sentiment analysis, and information extraction. Their ability to learn contextual representations from large-scale corpora has made transformer-based architectures a dominant approach in modern language processing systems [7]. While traditional sentiment analysis is limited to binary polarities, multiclass emotion classification offers more nuanced affective insights crucial for understanding customer behavior [8]. However, implementing these complex emotional representations in Indonesian e-commerce requires a specialized adaptation strategy, which remains underexplored.

One of the challenges often faced by Large Language Models (LLMs) is learning a new language or task over time without disrupting their performance in previously learned tasks, which are often dominated by English-language data [9]. BERT stands for Bidirectional Encoder Representations from Transformer, a language learning model used in various Natural Language Processing (NLP) tasks, such as understanding words context in a sentence and sentiment analysis [10] [11]. Language models like IndoBERT and IndoBERTweet can be utilized in Indonesia to carry out text classification tasks, which include the detection of emotions within text [12]. IndoBERT is a pre-trained model that uses the BERT algorithm, specifically adapted for the Indonesian language [13].

In the Indonesian language context, several pre-trained language models have been developed to address the unique linguistic characteristics of Indonesian text. One of the most widely used resources is the IndoNLU benchmark, which provides datasets and pre-trained models such as IndoBERT for various Natural Language Understanding tasks [14]. These models have enabled researchers to perform downstream tasks including sentiment analysis, text classification, and emotion detection more effectively in Indonesian. Continual fine-tuning, or CFT, is a step where an LLM is gradually modified to adapt to various tasks that have data variations and change over time [15]. The development of LLM applications for product review analysis focuses on the characteristics of the Indonesian language. This research uses English datasets with GPT and LLaMA models [16]. Conversely, our study seeks to develop a solution that is more culturally pertinent to the Indonesian e-commerce market. This emphasis on the local language is a distinct advantage, considering the inherent complexity and uniqueness of emotional expression in Indonesian.

Although transformer-based models have been widely applied for sentiment analysis in

Indonesian text, studies that explore their adaptation for multiclass emotion classification remain relatively limited. In particular, the application of continual fine-tuning to transfer knowledge from sentiment-oriented models to more complex emotion classification tasks has not been extensively investigated in Indonesian NLP.

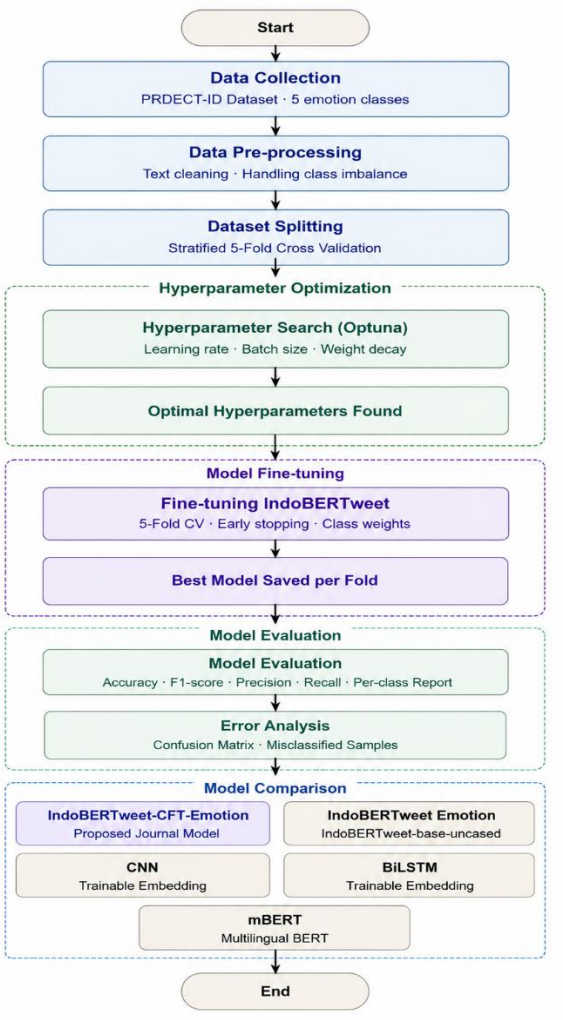
While prior studies have utilized CFT for cross-lingual adaptability [17], this research focuses on intra-language task adaptation. In contrast, our study specifically investigates the efficacy of CFT for enhancing a model's performance in a single language (Indonesian) for the more specific task of emotion classification. Our research also complements the qualitative study by Khamaludin et al. (2021) on the influence of marketing using social media on the marketing performance of Indonesian MSMEs [18]. The primary research gap lies in the difficulty of distinguishing high-arousal emotions with overlapping lexical markers in informal Indonesian. This study addresses this urgency by providing a precise AI-based diagnostic tool for MSMEs to interpret customer dissatisfaction beyond mere 'negative' labels."

The novelty of this study lies in the strategic application of CFT to repurpose sentiment-oriented pre-trained weights for a five-class emotion classification (Happiness, Anger, Fear, Love, and Sadness) without requiring exhaustive retraining. Methodologically, this demonstrates an efficient pathway for model evolution in low-resource settings. Consequently, the primary objective of this research is to evaluate the IndoBERTweet-CFT framework's ability to maintain contextual knowledge while adapting to complex emotional nuances. The findings are expected to provide a dual contribution: advancing the conceptual framework of knowledge transfer in Indonesian NLP and offering a practical, high-accuracy diagnostic tool for MSMEs to refine their digital marketing strategies.

MATERIALS AND METHODS

Research stages starting from collecting data to analysis of results are presented in Figure 1. it can be explained in more detail that data was collected using the PRDECT-ID dataset. The dataset used will be processed through data cleaning and normalization in the data preprocessing stage. The next step is to convert review text data into token representations that can be understood by the model, known as tokenization. The training process uses CFT on the Aardiiiiy/indobertweet-base-Indonesian-sentiment-analysis model for sentiment analysis. The optimization process in the training strategy uses the AdamW optimizer. Model evaluation uses macro F1-score and confusion matrix to measure the accuracy of classification results.

The PRDECT-ID dataset consists of five emotion categories: happiness, angry, neutral, surprised, and sad. However, the distribution of emotion labels in the dataset is not perfectly balanced, where certain emotion classes may appear less frequently than others. Therefore, macro F1-score is used as one of the primary evaluation metrics because it provides a balanced evaluation across classes.



Source: (Research Results, 2025)

Figure 1. Research Flow Diagram

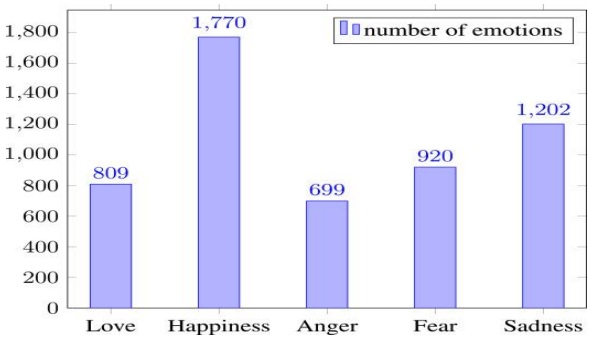
Figure 1 illustrates the proposed research methodology, which follows a deep learning pipeline structure. The methodology is designed as a direct replacement for previous, more generalized frameworks. Key stages include data acquisition from the PRDECT-ID dataset, comprehensive preprocessing, and tokenization tailored for the

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7729

IndoBERTweet model. The core methodology, highlighted by the dashed boundary, details the proposed Continual Fine-Tuning (CFT) mechanism, encompassing model initialization, adaptive training with the AdamW optimizer, and hyperparameter configuration. This structure emphasizes the specific contributions of this study to the localized Indonesian NLP context.

Data used in this research is taken from prdect-id dataset. The PRDECT-ID dataset contains five emotion categories: happiness, angry, neutral, surprised, and sad. However, the distribution of these emotion labels is not perfectly balanced, where certain classes such as *surprised* or *angry* may appear less frequently than others. This class imbalance may influence the classification performance and evaluation metrics. Dataset contains a collection of product reviews from Tokopedia platform Indonesian language that have been annotated for emotion and sentiment classification tasks. Figure 2 is a emotions distribution graph in prdect-id dataset.



Source: (Research Results, 2025)

Figure 2. Emotions Distribution in the Prdect-ID Dataset.

Dataset be fraught with 5400 reviews containing various attributes including product category, product name, seller location, price, sales total, number of reviews, sentiment labels (positive/negative) and emotions (Love, Happiness, anger, Fear, Sadness). Data was sourced from public reviews, adhering strictly to privacy policies and ensuring that no personally identifiable information is included. The emotion distribution in this dataset remains imbalanced, with a prevalence of happiness and sad emotions, whereas anger and love are less common. This imbalance reflects the emotional dynamics of consumers and challenges the process of training classification models. Nevertheless, prdect-id remains relevant for supporting emotion analysis models development, responsive chatbots, and various other NLP tasks such as opinion mining and text summarization within the context of Indonesian e-commerce.

The dataset utilized in this study was pre-cleaned and ready for use, therefore no additional text cleaning or normalization was necessary. Preprocessing stage performed was limited to the tokenization process, which is converting the text review data into a token representation that can be understood by a model. Tokenization was performed using a tokenizer from IndoBERTweet model (Aardiiiiy/ indobertweet-base-Indonesian-sentiment-analysis) which was specifically trained on Indonesian language texts from social media and digital platforms. This tokenization process uses padding and truncation with a maximum length of 128 tokens so that input data is uniform and efficient in training. Additionally, the textual emotion labels were converted into numerical format through a label encoder to simplify the model's classification. All preprocessing stages are carried out by utilizing a Huggingface Transformers and Datasets libraries with the result data is ready to be used for model training. The base model used in this study is IndoBERTweet, a transformer model based on the BERT architecture that has been specially trained on Indonesian language data from social media and online forums. This model has an edge in processing the unique informal language and sentence constructions of Indonesian users, making it a perfect fit for classifying product reviews, which tend to be spontaneous and expressive. A training process used is CFT which has previously been trained for sentiment analysis. The final classification layer was modified to accommodate multiclass emotion classification, with five distinct classes: Love, Happiness, anger, Fear, Sadness. The last layer uses a dense layer with a softmax activation function to mapping a token representation into probability distributions across five emotion classes.

This research employs a CFT approach, which involves further training IndoBERTweet with the prdect-id dataset. This strategy allows the model to understand complex emotion labels without losing the initial knowledge from pre-training, providing training efficiency by adapting to the new task domain without training from scratch. To ensure optimal model configuration, hyperparameter optimization was conducted using Optuna. Instead of relying on manual selection, we performed a systematic search across 5 trials to identify the best parameters. The optimization focused on three key hyperparameters: the learning rate (searched via a log-uniform distribution between 1e−5 and 5e−5), batch size (categorical selection of 8, 16, or 32), and weight decay (ranging from 0.01 to 0.1). This automated approach ensured

that the final model utilized the most effective combination for the dataset.

To address potential class imbalance and ensure unbiased evaluation, the dataset was partitioned using Stratified K-Fold Cross-Validation. This technique maintains the proportional distribution of emotion labels across both training and validation sets, preventing bias towards dominant classes during the evaluation phase. The model was trained using the AdamW optimizer. We employed an early stopping mechanism based on validation accuracy to prevent overfitting, ensuring the model was continuously adjusted based on the lowest validation loss.

Model performance was evaluated using a comprehensive set of metrics: Accuracy, Weighted F1-score, Precision, and Recall. While accuracy provides a general overview of correct predictions, the Weighted F1-score was specifically prioritized to address class imbalance. This metric calculates the harmonic mean of precision and recall, weighted by the number of true instances for each class, offering a robust measure for skewed datasets. In addition, a confusion matrix was utilized to analyze performance per class. This analysis is particularly valuable in emotion classification, where emotions like 'sadness' and 'fear' often share similar linguistic cues, necessitating a deeper evaluation beyond global accuracy to identify frequently confused classes. This study utilizes Huggingface Transformers as the primary architecture for modeling and training IndoBERTweet. The library provides an efficient interface for loading models, tokenizers, and training pipelines. The backend implementation was built using PyTorch, enabling tensor manipulation, GPU acceleration, and flexible debugging. All experiments were conducted on the Google Colab platform, which provides free GPU access and supports direct integration with Huggingface and PyTorch, creating a cost-efficient environment for rapid model training and evaluation.

RESULTS AND DISCUSSION

This section presents the comprehensive evaluation results of the IndoBERTweet-based continual fine-tuning (CFT) model for multiclass emotion classification on the Indonesian PRDECT-ID dataset. The analysis is structured into two main phases: Phase 1, dedicated to hyperparameter optimization using Optuna to identify the best configuration for the CFT process, and Phase 2, which details the final model performance evaluated via Stratified 5-Fold Cross-Validation. We discuss the model's performance using key metrics Accuracy, Weighted F1-score, Precision, and Recall and provide a misclassification analysis to identify specific confusion patterns between emotion classes. Finally, the results are benchmarked against previous studies and baseline models to validate the contribution of the proposed approach. The first phase of the experiment focused on identifying the optimal configuration using the Optuna framework. Given the constraints of the PRDECT-ID dataset's linguistic complexity, a Bayesian search through 5 trials was conducted on Fold-1 to minimize validation loss.

Table 1. Hyperparameter Configuration For The Continual Fine-Tuning Process

Parameter	Search Range	Value
Optimizer	-	AdamW
Learning Rate	1×10^{-5} – 5×10^{-5} (log-unif)	2.17×10^{-5} (optuna)
Weight Decay	0.01 – 0.10	0.0498 (Optuna)
Batch Size	{8, 16, 32}	8 (Optimized via Optuna)
Max Epochs	-	5
Early Stopping	-	Patience = 2
Cross-validation	-	Stratified 5-Fold

Source: (Research Results, 2025)

The search results indicate that a relatively low learning rate (2.17×10^{-5}) and a small batch size of 8 were optimal for the CFT process. The small batch size suggests that more frequent gradient updates were necessary to capture the nuances of the five emotion classes, while the early stopping patience of 2 served as a crucial safeguard against overfitting, especially given that the search process found 10 epochs to be the upper limit for stable convergence.

The PRDECT-ID dataset exhibits a notable class imbalance, as detailed in Table 2. This imbalance can lead to a "majority class bias" where the model over-predicts common emotions like "Happiness" while neglecting minority classes like "Anger."

Table 2. - Class Distribution and Weights PRDECT-ID Dataset

Emotion	Samples	(%)	Class Weight
Happiness	1770	32.8%	0.610
Sadness	1202	22.3%	0.899
Fear	920	17.0%	1.174
Love	809	15.0%	1.335
Anger	699	12.9%	1.545
Total	5400	100%	-

Source: (Research Results, 2025)

To mitigate this, we implemented cost-sensitive learning by calculating inverse frequency class weights. The "Anger" class received the

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7729

highest weight (1.545), effectively penalizing the model more heavily for misclassifying this minority class. This strategy ensures that the model remains sensitive to critical negative feedback in product reviews, which is often more valuable for sentiment analysis than generic positive feedback. To ensure the generalizability of the IndoBERTweet-CFT model, we conducted a Stratified 5-Fold Cross-Validation. This approach provides a more rigorous estimate of performance by ensuring each fold maintains the same class distribution as the original dataset, thereby reducing bias associated with a single train-test split. The consolidated results are presented in Table 3.

Table 3. 5-Fold Cross Validation Results

Fold	Accuracy	F1 Weighted	F1 Macro	Precision	Recall
Mean	0.7137 ±0,026	0.7138 ±0,025	0,6881 ±0,026	0,7185 ±0,020	0,7137
Std	6	2	1	8	±0,0266

Source: (Research Results, 2025)

The model achieved a mean accuracy of 0,7137 and a Weighted F1-score of 0,7138. The minimal discrepancy between Accuracy and Weighted F1-score suggests that the cost-sensitive learning approach (class weighting) effectively balanced the model's predictive power across the imbalanced classes. However, the Macro F1-score (0.6881) is notably lower than the Weighted F1.

This gap indicates that while the model performs exceptionally well on majority classes like "Happiness," it still faces significant challenges in maintaining identical precision and recall for minority classes such as "Anger" and "Fear." A deeper dive into the classification report presented in Table 4 reveals the specific strengths and vulnerabilities of the IndoBERTweet-CFT-Emotion model across the emotional spectrum. The results indicate that model performance is not only influenced by the volume of training data (support) but also by the linguistic distinctiveness of each emotion within the Indonesian e-commerce context.

Table 4. Overall Per-Class Classification Report (All Folds)

Class	Precision	Recall	F1 Score	Support
Anger	0.6449	0.6080	0.6259	699
Fear	0.5559	0.5783	0.5669	920
Happiness	0.8557	0.8311	0.8432	1770
Love	0.6914	0.7590	0.7236	809
Sadness	0.6899	0.6755	0.6826	1202

Source: (Research Results, 2025)

The "Happiness" class significantly outperformed all other categories. This superior performance is a result of lexical redundancy and high support volume. With 1770 samples, the model achieved a high Precision (0,8557), indicating a low false-positive rate. In Indonesian product reviews, satisfaction is often conveyed through standardized, high-frequency lexicons such as "mantap", "sesuai", "cepat", and "puas". The model's ability to recognize these explicit positive markers allows it to distinguish Happiness from negative emotions with high reliability. In contrast, "Fear" proved to be the most challenging class for the model, yielding the lowest Precision (0.5559) and F1-score. Unlike Happiness, fear in an e-commerce setting is rarely expressed through primary emotion words. Instead, it is often contextually embedded in concerns about product authenticity, warranty issues, or shipping damage (e.g., "takut barangnya palsu" or "khawatir pecah"). The lower precision suggests that the model frequently misidentifies other negative sentiments as Fear, likely due to the overlapping semantic space between anxiety and general disappointment.

The "Love" class exhibited a notably high Recall (0,7590) but a relatively lower Precision (0,6914). This disparity indicates that while the model is highly capable of capturing "Love" instances, it often struggles with intensity-based boundaries. In many reviews, the distinction between being "Happiness" (satisfied with a product) and "Love" (brand loyalty or extreme appreciation) is subtle. This overlap results in the model occasionally over-predicting Love for reviews that might only represent high levels of Happiness, a phenomenon known as valence-level clustering. "Anger" (F1: 0,6259) and "Sadness" (F1: 0,6826) represent the model's mid-tier performance. "Anger" suffered from the lowest Recall (0,6080) among all classes, suggesting that many angry reviews are being "softened" by the model and misclassified as Sadness or Fear. This is often due to the informal syntax used in Indonesian complaints; users may express frustration through capitalization or repetitive punctuation which, while clear to humans, may be mapped by the model to general negativity rather than specific "Anger." "Sadness," which often correlates with "disappointment" (kecewa) in reviews, remains more stable because the term "kecewa" acts as a strong, relatively unambiguous anchor for the model.

To further investigate the model's decision-making process and identify problematic semantic boundaries, we analyzed the Confusion Matrix presented in Table 5. This matrix provides a

granular view of the "misclassification paths," revealing how the model perceives the relationship between different Indonesian emotional expressions in an e-commerce context.

Table 5. Overall Confusion Matrix

Class	Anger	Fear	Happiness	Love	Sadness
Anger	425	148	5	9	112
Fear	133	532	22	2	231
Happiness	0	20	1471	261	18
Love	1	0	190	614	4
Sadness	100	257	31	2	812

Source: (Research Results, 2025)

The most significant confusion pattern observed is the bidirectional overlap between Fear and Sadness. The model misclassified 231 instances of Fear as Sadness and 257 instances of Sadness as Fear. In the linguistic landscape of Indonesian product reviews, these two emotions often share a convergent semantic structure. When a customer receives a defective item, the expression of disappointment ("kecewa") which is typically mapped to Sadness is often inextricably linked with an underlying anxiety about financial loss or the hassle of a refund ("takut rugi", "khawatir tidak bisa retur") which is mapped to Fear. This creates a "gray area" where the transformer's attention mechanism struggles to isolate the primary emotion without explicit, high-intensity lexicons.

A prominent confusion also occurs within the positive polarity, where 261 instances of Happiness were predicted as Love, and 190 instances of Love were predicted as Happiness. Unlike the negative cluster, this confusion is primarily driven by intensity scaling rather than semantic ambiguity. Reviews labeled as "Love" often contain high-praise keywords like "langganan" (subscriber/loyalty) or "cinta banget", but they frequently appear alongside standard "Happiness" markers like "barang bagus" and "pengiriman cepat". Since both classes reside in the same positive valence region, the IndoBERTweet-CFT model occasionally fails to distinguish between "satisfied compliance" (Happiness) and "enthusiastic loyalty" (Love), suggesting that these categories are perceived as a continuum rather than discrete classes.

The "Anger" class exhibits a unique error pattern where it is frequently misclassified as either Fear (148 cases) or Sadness (112 cases). This suggests a "sentiment softening" effect in the model's predictions. Angry reviews often utilize informal markers such as all-caps text, excessive exclamation marks, and harsh slang. While the IndoBERTweet pre-training includes Twitter data rich in such features, the specific domain of product reviews may use "softer" anger (e.g., sarcastic politeness) that the model interprets as Sadness (disappointment) or Fear (concern over a faulty product). The 133 cases where Fear was misclassified as Anger further emphasize that high-arousal negative emotions in the PRDECT-ID dataset share a very narrow decision boundary.

One of the most impressive findings from Table 5 is the model's robustness against cross-polarity errors. There were zero instances of "Happiness" being misclassified as "Anger," and only a negligible number of positive reviews were predicted as negative (e.g., 18 cases of Happiness as Sadness). This indicates that while the model may struggle with fine-grained intra-polarity nuances (distinguishing between two negative or two positive emotions), it has mastered the fundamental distinction between positive and negative sentiment, which is critical for maintaining overall accuracy in practical e-commerce monitoring applications.

The final evaluation phase involves benchmarking the proposed IndoBERTweet-CFT-Emotion model against a diverse set of baseline architectures. This comparison includes traditional deep learning models (CNN and BiLSTM), a global transformer baseline (mBERT), and the vanilla domain-specific transformer (IndoBERTweet-base-uncased-emotion-recognition). The results, summarized in Table 6, provide clear evidence of the performance gains achieved through the continual fine-tuning (CFT) approach.

Table 6. Model Performance Comparison

Model	Accuracy	Weighted F1
IndoBERTweet-CFT-Emotion	0,81574	0.81184
indobertweet-base-uncased-emotion-recognition	0,79444	0.79139
CNN	0,60555	0.60397
BiLSTM	0,58055	0.57885
mBERT (Multilingual)	0,22037	0.20495

Source: (Research Results, 2025)

The results demonstrate a massive performance gap between the IndoBERTweet lineage and the mBERT baseline. Achieving only 22.03% accuracy, mBERT failed to provide a viable classification for Indonesian product reviews. This "collapse" is primarily due to tokenization mismatch and the curse of multilinguality. mBERT's sub-word vocabulary is shared across over 100 languages, which often results in Indonesian slang, abbreviations (e.g., "brng", "sdh", "gak"), and informal e-commerce terms being fragmented into semantically hollow tokens. In contrast, IndoBERTweet, which was pre-trained on millions of Indonesian tweets, preserves the semantic integrity of informal Indonesian syntax.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7729

While CNN (60,55%) and BiLSTM (58,05%) performed significantly better than mBERT, they were unable to match the contextual understanding of the transformer-based models. Traditional architectures rely on localized feature extraction (CNN) or sequential processing (BiLSTM), which struggle to capture the long-range dependencies and contextual nuances present in multi-sentence product reviews. Furthermore, without the benefit of large-scale pre-training, these models are more susceptible to the noise and high variance of the PRDECT-ID dataset.

The core contribution of this research is validated by the 2,13% performance gain of the IndoBERTweet-CFT-Emotion model over the indobertweet-base-uncased-emotion (0,7944). In the competitive landscape of NLP benchmarks, a 2% improvement is statistically significant. This improvement indicates that Continual Fine-Tuning serves as an essential "adaptation layer." While the base IndoBERTweet model is highly proficient in general social media Indonesian, product reviews contain a unique distribution of emotional signals often revolving around logistics, packaging, and functional utility that differ from general social media discourse. The CFT approach successfully recalibrated the model's attention weights to these specific emotional nuances, effectively bridging the gap between general pre-training and specialized domain application.

The achievement of 81,57% accuracy in a multiclass (5-class) environment is a robust result for automated sentiment monitoring systems. For e-commerce stakeholders, this high level of precision particularly in identifying negative emotions enables more effective automated filtering of critical customer complaints. The consistency between accuracy and Weighted F1-score across the benchmarks confirms that the proposed CFT strategy provides a stable and reliable framework for handling the linguistic messiness of real-world Indonesian consumer feedback

Limitations

This study has several limitations. First, class imbalance in PRDECT-ID persists despite using class-weighted loss; minority classes like Fear and Anger remain difficult, with F1-scores of 0.5669 and 0.6259. Second, model comparison includes IndoBERTweet, mBERT, BiLSTM, and CNN, but transformer models were used without fine-tuning, limiting performance insights. Third, hyperparameter tuning with Optuna used only five trials, so better configurations may exist. Finally, although evaluation is comprehensive, 28.6% of samples were misclassified, mainly between similar emotions (e.g., Fear–Sadness), indicating ongoing challenges in distinguishing subtle linguistic expressions

CONCLUSION

This study provides robust empirical evidence that Continual Fine-Tuning (CFT) serves as a highly effective and efficient adaptation strategy for transformer-based models within the Indonesian NLP landscape. By strategically repurposing the pre-trained IndoBERTweet model, this approach achieved a superior performance benchmark with an accuracy of 0,8157 and a weighted F1-score of 0,8118. The primary scientific contribution lies in the methodology's ability to leverage sentiment-based contextual knowledge and successfully transfer it to a more complex five-class emotion task. This demonstrates that "knowledge repurposing" can bypass the need for exhaustive retraining from scratch, providing a significant advantage in domain-specific tasks.

From a theoretical perspective, these findings offer a new paradigm for developing local-language Large Language Models (LLMs), particularly in low-resource scenarios. This study proves that specialized emotional knowledge can be "layered" onto existing models, preserving semantic depth while expanding functional breadth. However, certain academic limitations must be acknowledged. The PRDECT-ID dataset's scale (5400 samples) and the inherent subjectivity in affective labeling remains a challenge for broader generalization. Specifically, the high semantic overlap between categories such as "Fear" and "Sadness" reflects a fundamental difficulty in mapping the nuanced emotional lexicon of Indonesian consumer reviews, where disappointment and anxiety often converge.

In conclusion, while offering immediate practical utility for e-commerce sentiment monitoring for MSMEs, this research contributes to the broader NLP literature by validating CFT as a robust framework for local-language model evolution. Practical implications of this study suggest that MSMEs can integrate this model into customer service dashboards to prioritize 'Anger' and 'Fear' labels, which require immediate intervention compared to general 'Sadness'. Future work should explore the integration of Easy Data Augmentation (EDA) to bolster the representation of minority classes and investigate the model's performance on cross-platform e-commerce data.

REFERENCE

[1] H. Alqam, M. Razzak, A. Al-Busaidi, and S. Al-Riyami, "Conceptualizing Digital Readiness, Strategic Foresight, and Strategic Flexibility as Drivers of Digitalization," Int. J. Informatics Vis., vol. 8, no. 2, pp. 938–947, May 2024, doi: 10.62527/joiv.8.2.2230.

[2] R. Kaur, R. Singh, A. Gehlot, N. Priyadarshi, and B. Twala, "Marketing Strategies 4.0: Recent Trends and Technologies in Marketing," *Sustain.*, vol. 14, no. 24, pp. 1–17, 2022, doi: 10.3390/su142416356.

[3] A. Romadhony, S. Al Faraby, R. Rismala, U. N. Wisesti, and A. Arifianto, "Sentiment Analysis on a Large Indonesian Product Review Dataset," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 1, pp. 167–178, 2024, doi: 10.20473/jisebi.10.1.167-178.

[4] Y. Wang, E. W. T. Ngai, and K. Li, "The effect of review content richness on product review helpfulness: The moderating role of rating inconsistency," *Electron. Commer. Res. Appl.*, vol. 61, p. 4223, 2023, doi: 10.1016/j.elerap.2023.101290.

[5] M. Alzate, M. Arce-Urriza, and J. Cebollada, "Online reviews and product sales: The role of review visibility," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 4, pp. 638–669, 2021, doi: 10.3390/jtaer16040038.

[6] L. A. Huwaida *et al.*, "Generation Z and Indonesian Social Commerce: Unraveling key drivers of their shopping decisions," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 2, p. 100256, 2024, doi: 10.1016/j.joitmc.2024.100256.

[7] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, 2024, doi: 10.1016/j.jksuci.2024.102048.

[8] Q. Zhao, Y. Xia, Y. Long, G. Xu, and J. Wang, "Leveraging sensory knowledge into Text-to-Text Transfer Transformer for enhanced emotion analysis," *Inf. Process. Manag.*, vol. 62, no. 1, 2025, doi: 10.1016/j.ipm.2024.103876.

[9] E. Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," *Comput. Educ. Artif. Intell.*, vol. 6, no. December 2023, p. 100210, 2024, doi: 10.1016/j.caeai.2024.100210.

[10] N. C. Mei, S. Tiun, and G. Sastria, "Multi-Label Aspect-Sentiment Classification on Indonesian Cosmetic Product Reviews with IndoBERT Model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 11, pp. 712–720, 2024, doi: 10.14569/IJACSA.2024.0151168.

[11] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, "Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 11, no. 1, pp. 30–40, 2025, doi: 10.20473/jisebi.11.1.30-40.

[12] A. Alamsyah and Y. Sagama, "Empowering Indonesian internet users: An approach to counter online toxicity and enhance digital well-being," *Intell. Syst. with Appl.*, vol. 22, no. August 2023, p. 200394, 2024, doi: 10.1016/j.iswa.2024.200394.

[13] F. Baharuddin and M. F. Naufal, "Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 9, no. 2, pp. 253–263, 2023, doi: 10.20473/jisebi.9.2.253-263.

[14] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," *EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 8875–8898, 2021, doi: 10.18653/v1/2021.emnlp-main.699.

[15] Z. Zhou, J. Y. Shin, S. R. Gurudu, M. B. Gotway, and J. Liang, "Active, continual fine tuning of convolutional neural networks for reducing annotation effortsL Active continual fine tuningcnnrae," *Med. Image Anal.*, vol. 71, pp. 1–38, 2021, doi: 10.1016/j.media.2021.101997.

[16] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation," *Nat. Lang. Process. J.*, vol. 6, p. 100056, Mar. 2024, doi: 10.1016/J.NLP.2024.100056.

[17] D. Aggarwal, S. Damle, N. Goyal, S. Lokam, and S. Sitaram, "Exploring Continual Fine-Tuning for Enhancing Language Ability in Large Language Model," 2024, [Online]. Available: https://arxiv.org/abs/2410.16006.

[18] Khamaludin *et al.*, "The influence of social media marketing, product innovation and market orientation on indonesian smes marketing performance," *Int. J. Data Netw. Sci.*, vol. 6, no. 1, pp. 9–16, 2021, doi: 10.5267/J.IJDNS.2021.11.002.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri