Vol. 8, No. 3, June 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Class-Level Behavior Analysis under Metric Disagreement in Imbalanced Multi-Label Indonesian Emotion Classification

Jahda Rusti Putri[1], Ermatita[2], Abdiansah[3]

[1,2,3] Magister of Computer Science, Faculty of Computer Science, Sriwijaya University, Indonesia

Received:
October 13, 2025
Revised:
May 17, 2026
Accepted:
June 6, 2026
Published:
June 25, 2026

Corresponding Author:
Author Name*:
Ermatita
Email*:
Ermatita@unsri.ac.id

DOI:
10.63158/journalisi.v8i3.1664

Abstract. This study aims to analyze class-level model behavior under metric disagreement in imbalanced multi-label Indonesian emotion classification, using the divergence between Macro F1 and Micro F1 as a diagnostic signal rather than a mere performance indicator. A machine-translated Indonesian version of the GoEmotions dataset, comprising approximately 58,000 samples across 28 fine-grained emotion categories, is used as the experimental setting. The translated dataset was not manually revalidated, and findings are scoped to this translated GoEmotions setting. Two transformer-based models are evaluated: IndoBERT, a monolingual Indonesian model, and DistilBERT, a multilingual model, both fine-tuned with class-specific threshold optimization. The results reveal opposing divergence patterns: IndoBERT achieves higher Micro F1 than Macro F1, indicating performance concentrated on high-frequency classes, while DistilBERT exhibits the reverse pattern, suggesting broader but less precise label activation. Per-class analysis further shows that most minority classes consistently fall into unstable or non-functional performance regimes across both models. This study concludes that aggregate metrics alone are insufficient for evaluating model behavior in imbalanced multi-label settings. A behavior-oriented interpretation framework for Macro–Micro F1 divergence and a regime-based class reliability categorization are proposed to support more structured and informative evaluation practices.

Keywords: evaluation metrics, metric divergence, multi-label classification, class imbalance, emotion classification.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Evaluation metrics are commonly used to summarize model performance in machine learning [1]. In many cases, these metrics are treated as reliable indicators of what a model can achieve. This assumption is generally reasonable when datasets are balanced and consist of a limited number of classes [2]. However, in imbalanced multi-label classification, particularly in emotion recognition tasks, aggregate metrics may not fully represent how models perform across different classes [3], [4]. Macro F1 assigns equal weight to all labels regardless of class frequency, making it sensitive to minority-class performance, while Micro F1 aggregates predictions across all instances, making it more influenced by frequent classes. Under such conditions, aggregate scores may conceal substantial variation in class-level performance, and the divergence between these two metrics may provide additional insight into model behavior beyond overall performance estimates[5].

The development of emotion classification in natural language processing has gradually evolved from binary sentiment analysis toward more fine-grained representations of human affect [6], [7]. Early approaches focused primarily on distinguishing positive and negative polarity, which proved insufficient for applications requiring deeper emotional understanding, such as mental health monitoring, customer feedback analysis, and content moderation [8], [9]. This limitation encouraged the adoption of categorical emotion models, including the six-emotion framework proposed by Ekman [10], as well as more detailed taxonomies capable of capturing the diversity of emotional expression in natural language.

The emergence of large-scale annotated datasets further accelerated research in this area. The GoEmotions dataset [11] provides one of the most comprehensive benchmarks for multi-label emotion classification, consisting of 28 emotion categories annotated from social media text. Using transformer-based models, previous studies commonly reported Macro F1 scores in the range of 0.46–0.50 [11]. Although these results demonstrate the feasibility of fine-grained emotion classification, evaluation is generally reported in aggregate form, with limited discussion regarding variation across individual classes. This gap is particularly consequential when class distributions are highly

imbalanced, as similar aggregate scores can arise from fundamentally different prediction strategies across the label space [3], [12].

These challenges become more pronounced in low-resource settings such as Indonesian NLP. Although IndoBERT and related language resources have improved performance across various Indonesian NLP tasks [13], [14], [15], [16], [17], existing evaluations are commonly conducted on relatively balanced datasets or simpler classification settings. In this study, the GoEmotions dataset is translated into Indonesian to enable large-scale experimentation in a low-resource, fine-grained emotion classification context, as no existing Indonesian emotion dataset offers comparable scale and label granularity across 28 categories.

The use of translated datasets introduces additional complexity, since cross-lingual transfer may not fully preserve culturally specific emotional expressions, particularly for fine-grained emotion categories [18], [19]. This issue becomes especially relevant for low-frequency classes, where both limited data availability and translation effects may influence prediction reliability. Recent studies in NLP have increasingly emphasized behavior-oriented evaluation approaches that focus not only on aggregate performance, but also on how models distribute predictions across classes [15], [19]. These studies suggest that models with similar overall scores may exhibit substantially different behavior when examined at a finer level of analysis.

The core problem addressed in this study is the interpretive ambiguity of aggregate evaluation metrics in imbalanced multi-label classification. When a model reports substantially different Macro F1 and Micro F1 values, it is not immediately clear whether this gap reflects strong performance on frequent classes, poor performance on minority classes, or a fundamental difference in prediction strategy. Without class-level analysis, this ambiguity cannot be resolved, and evaluation results may be misleading [20]. Furthermore, practitioners relying solely on aggregate metrics may select or deploy models based on incomplete information about their actual behavior across the full label space, with potentially significant consequences for real-world applications such as Indonesian emotion-aware systems [21].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

To frame this analysis precisely, three related but distinct concepts are distinguished in this study. Metric disagreement refers to the phenomenon where Macro F1 and Micro F1 produce different numerical outcomes for the same model. Metric divergence specifically denotes the direction and magnitude of the gap between these two metrics, which carries diagnostic information about how predictions are distributed across classes. Model behavior refers to the distributional pattern of predictions across the label space as inferred from these metrics and from per-class performance analysis.

To address these issues, this study adopts a behavior-oriented analysis framework. The primary contribution of this work is not a new classification model or training strategy, but an evaluation interpretation framework. Specifically, this study contributes: (1) a structured interpretation of Macro–Micro F1 divergence as a diagnostic indicator of how prediction capacity is distributed across the label space, rather than merely a reporting inconsistency; and (2) a regime-based class reliability categorization that groups per-class performance into high-confidence, unstable, and non-functional regions, providing a clearer and more structured perspective on model reliability in imbalanced multi-label classification settings.

Guided by this framework, the study is structured around three research questions: (RQ1) How does the direction and magnitude of Macro-Micro F1 divergence reflect differences in prediction behavior between a monolingual and a multilingual transformer model in imbalanced multi-label emotion classification? (RQ2) How are per-class performance levels distributed across distinct reliability regimes? (RQ3) What practical implications does behavior-oriented metric interpretation offer a model selection in Indonesia emotion classification applications?

This study therefore examines how the direction and magnitude of Macro–Micro F1 divergence can be interpreted as a behavioral signal in imbalanced multi-label emotion classification, using IndoBERT and DistilBERT as contrasting cases. The analysis proceeds at the class level to identify how prediction behavior differs across emotion categories with varying data availability, and how class-level performance can be systematically structured to support more reliable and interpretable model evaluation.

2. METHODS

The overall research methodology adopted in this study is illustrated in Figure 1. The proposed framework is designed to analyze how evaluation metrics reflect model behavior under imbalanced multi-label conditions, rather than to optimize model performance. The workflow consists of five main stages:

1) Data Preparation, including the use of a translated version of the GoEmotions dataset and stratified splitting into training, validation, and test sets.
2) Model Training, using two transformer-based architectures, IndoBERT and DistilBERT, for multi-label emotion classification.
3) Threshold Optimization, where class-specific decision thresholds are determined using validation data to improve F1-score under imbalanced conditions.
4) Behavior-Oriented Analysis, consisting of metric comparison, per-class performance analysis, and precision-recall examination to interpret model behavior.
5) Regime Categorization, where class-level performance is grouped into high-confidence, unstable, and non-functional regions to provide a structured interpretation of reliability across the label space.

2.1. Dataset Preparation

This study uses a translated version of the GoEmotions dataset [11], which consists of approximately 58,000 English Reddit comments annotated with 28 emotion categories in a multi-label setting. The dataset is translated into Indonesian using the Google Translate API to enable experimentation in a low-resource context. The translation was performed programmatically using the Google Translate API without manual quality verification. No manual revalidation or native-speaker review was conducted. Consequently, translated texts were not systematically inspected for broken, incomplete, or grammatically unnatural outputs. This approach was chosen for scalability given the dataset size. It is acknowledged that machine translation may not fully preserve the semantic nuance of fine-grained emotion categories, particularly for culturally specific expressions. Emotional label assignments from the original English annotation are assumed to remain valid after translation; however, this assumption constitutes a limitation of the study, as

some fine-grained distinctions may be altered by the translation process. This limitation is discussed further in Section 3.7.

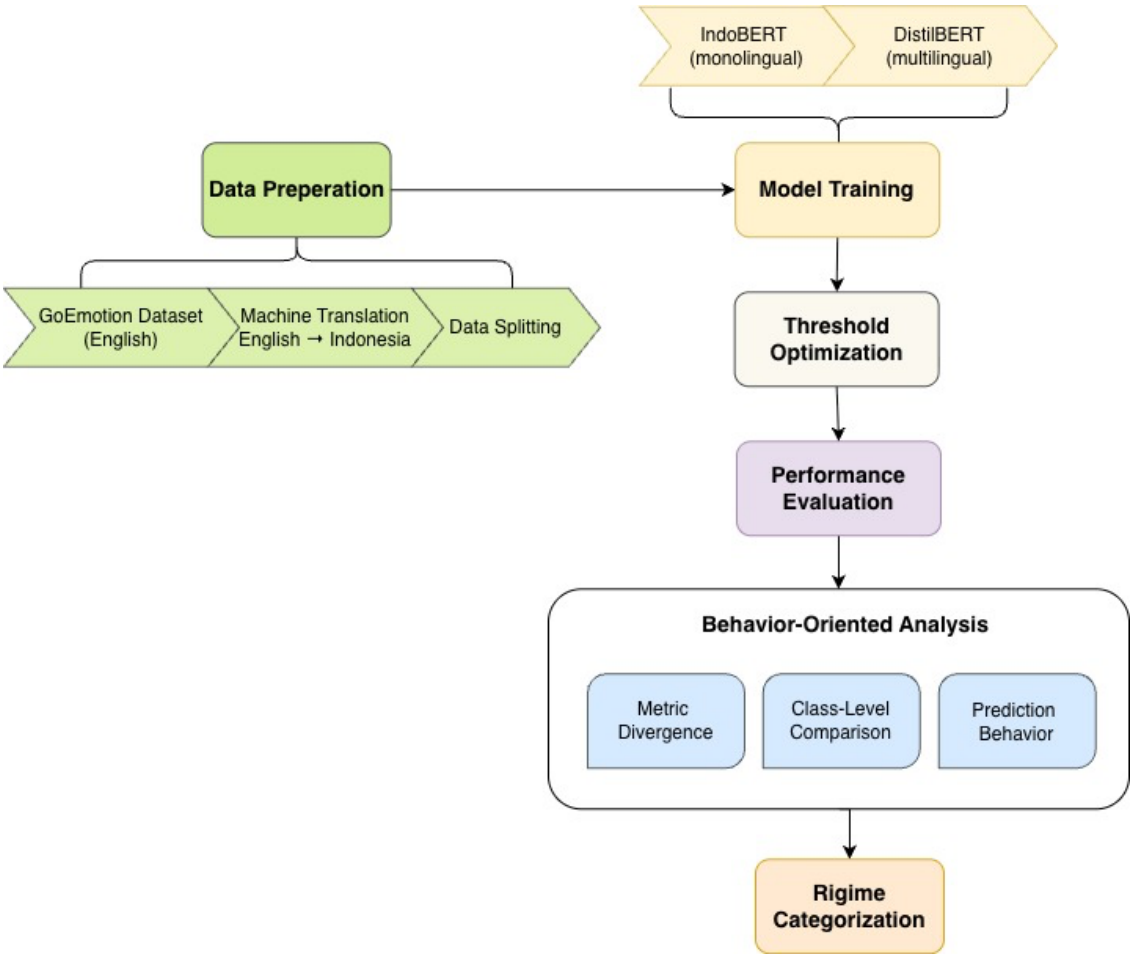

Figure 1. Research workflow for behavior-oriented analysis in imbalanced multi-label emotion classification.

The resulting dataset exhibits a highly imbalanced distribution, where a small number of emotion categories dominate the data while many others appear only infrequently. This distribution reflects real-world patterns of emotional expression in social media [11], [22] and is therefore preserved without modification. No resampling or data augmentation is applied, as the objective of this study is to analyze model behavior under natural data conditions. To ensure consistent evaluation, the dataset is divided into training, validation, and test sets using an 80:10:10 split with stratified sampling. This preserves the relative class distribution across all subsets. Table 1 presents the full class distribution across all 28 emotion categories. The distribution confirms the highly imbalanced nature of the

dataset, with neutral accounting for the largest proportion and several low-frequency categories such as grief, pride, and nervousness containing fewer than 200 instances in the test set.

Table 1. Class Distribution of the Translated GoEmotions Dataset Across All 28 Emotion Categories

Emotion Label	Train	Val	Test	Total	Regime (approx.)
neutral	9,530	1,191	1,192	11,913	High-confidence
admiration	3,756	470	469	4,695	High-confidence
amusement	2,511	314	313	3,138	High-confidence
approval	2,329	291	291	2,911	High-confidence
caring	1,215	152	152	1,519	Unstable
curiosity	1,205	150	151	1,506	Unstable
desire	755	94	95	944	Unstable
disapproval	1,844	231	230	2,305	Unstable
disappointment	983	123	123	1,229	Unstable
disgust	755	94	95	944	Unstable
embarrassment	252	31	32	315	Non-functional
excitement	1,091	137	136	1,364	Unstable
fear	367	46	46	459	Unstable
gratitude	2,980	372	372	3,724	High-confidence
grief	104	13	13	130	Non-functional
joy	1,355	169	170	1,694	Unstable
love	1,217	152	152	1,521	High-confidence
nervousness	234	29	30	293	Non-functional
optimism	1,177	147	147	1,471	Unstable
pride	131	16	17	164	Non-functional
realization	952	119	119	1,190	Unstable
relief	214	27	27	268	Non-functional
remorse	538	67	67	672	Unstable
sadness	1,314	164	165	1,643	Unstable
surprise	1,188	149	148	1,485	Unstable
anger	1,392	174	174	1,740	Unstable
annoyance	2,985	373	373	3,731	Unstable
confusion	1,092	136	137	1,365	Unstable

To ensure consistent label distribution across splits, iterative stratification was applied a method designed specifically for multi-label datasets where standard single-label stratification is insufficient [23]. This approach iteratively assigns samples to splits to approximate the desired label proportions across all 28 classes. The dataset is divided into training (80%), validation (10%), and test (10%) sets. Neutral labels and multi-emotion samples were handled without special treatment, following the protocol of the original GoEmotions dataset. A fixed random seed of 42 was used for all splits and training procedures.

2.2. Model Configuration

Two transformer-based models are evaluated in this study. The first model is IndoBERT, using the pretrained checkpoint indobenchmark/indobert-base-p1, a monolingual BERT model trained on large-scale Indonesian corpora [14], [15]. The second model is DistilBERT, using the pretrained checkpoint distilbert-base-multilingual-cased, a multilingual transformer model designed to support multiple languages through knowledge distillation [21], [22]. These models were selected to represent two contrasting paradigms in Indonesian NLP: monolingual specialization and multilingual generalization. Both models were fine-tuned for multi-label classification using a shared output architecture consisting of a dropout layer followed by a sigmoid output layer with 28 units. Each unit corresponds to one emotion category and produces an independent probability estimate.

Input texts were processed using the tokenizer associated with each pretrained checkpoint. Sequences were tokenized and adjusted to a maximum length of 128 tokens through padding and truncation. To ensure reproducibility, all experiments were conducted using a fixed random seed of 42. The experiments were performed using a single train–validation–test split without cross-validation, which is acknowledged as a limitation of this study. The complete training configuration, including hyperparameter settings and hardware specifications, is presented in Table 2.

Table 2. Training Configuration for Both Model

Parameter	IndoBERT	DistilBERT
Model checkpoint	indobenchmark/indobert-base-p1	distilbert-base-multilingual-cased
Language scope	Indonesian	Multilingual (104 languages)
Tokenizer	Checkpoint tokenizer	Checkpoint tokenizer

Parameter	IndoBERT	DistilBERT
Learning rate	2e-5	2e-5
Batch size	32	32
Optimizer	AdamW	AdamW
Number of epochs	5	5
Dropout rate	0.3	0.3
Maximum sequence length	128 tokens	128 tokens
Random seed	42	42
Hardware	NVIDIA T4 GPU (Google Colab)	NVIDIA T4 GPU (Google Colab)
Output layer	Dropout + Sigmoid (28 units)	Dropout + Sigmoid (28 units)
Software environment	Python 3.10, PyTorch 2.0.1, HuggingFace Transformers 4.35.0, scikit-learn 1.3.0, Numpy 1.24	Python 3.10, PyTorch 2.0.1, HuggingFace Transformers 4.35.0, scikit-learn 1.3.0, Numpy 1.24

2.3. Threshold Optimization

A fixed decision threshold of 0.5 is commonly used in multi-label classification but may not be suitable for imbalanced data [24]. Models trained under such conditions tend to underestimate prediction confidence for minority classes, leading to low recall. To address this, class-specific thresholds are optimized using validation data. A grid search over the range [0.1, 0.9] is performed, and the threshold that maximizes the F1-score for each class is selected. These thresholds are then applied to the test set for evaluation.

The resulting optimized thresholds are presented in Table 3. Considerable variation can be observed across emotion categories, particularly for IndoBERT, where several classes require substantially lower thresholds than the conventional value of 0.5. Such adjustments increase sensitivity toward minority classes that may otherwise be under-predicted. DistilBERT exhibits a narrower threshold distribution, suggesting a more uniform confidence calibration across emotion categories.

Table 3. Optimized Class-Specific Decision Threshold for IndoBERT and DistilBERT Across All 28 Emotion Classes

Emotion	IndoBERT	DistilBERT
Admiration	0.45	0.550
Amusement	0.45	0.600
Anger	0.30	0.550

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Emotion	IndoBERT	DistilBERT
Annoyance	0.15	0.550
Approval	0.15	0.550
Caring	0.10	0.550
Confusion	0.30	0.575
Curiosity	0.30	0.575
Desire	0.15	0.575
Disappointment	0.10	0.550
Disapproval	0.25	0.550
Disgust	0.35	0.550
Embarrassment	0.70	0.525
Excitement	0.25	0.575
Fear	0.20	0.550
Gratitude	0.75	0.625
Grief	0.15	0.500
Joy	0.15	0.550
Love	0.30	0.575
Nervousness	0.15	0.525
Optimism	0.20	0.575
Pride	0.55	0.525
Realization	0.25	0.550
Relief	0.20	0.500
Remorse	0.25	0.525
Sadness	0.25	0.575
Surprise	0.10	0.550
Neutral	0.20	0.550

Table 3 shows distinct threshold distributions between the two models. IndoBERT exhibits a wider threshold range (0.10–0.75), indicating substantial variation in confidence calibration across classes. Several minority classes, including Caring, Disappointment, and Surprise, require thresholds below 0.20, suggesting that higher sensitivity is needed to detect these less frequent labels. In contrast, DistilBERT produces a narrower threshold range (0.50–0.625), reflecting more consistent probability estimates across emotion categories. These differences indicate that the two models respond differently to class imbalance despite being trained on the same dataset.

It is important to note that threshold optimization is performed exclusively on the validation set. While this procedure improves class-level F1 performance under imbalanced conditions, it introduces a potential risk of threshold overfitting. The selected thresholds may partially reflect characteristics specific to the validation distribution and therefore may not generalize equally well to unseen datasets or alternative train–test splits. This limitation is acknowledged and further discussed in Section 3.7.

2.4. Performance Evaluation and Behavior-Oriented Analysis

Model performance is evaluated using Micro F1, Macro F1, Macro Precision, Macro Recall, Micro Precision, Micro Recall, per-class F1 scores, and Hamming Loss [25], [26]. Hamming Loss complements the F1-based metrics by measuring label-level prediction errors across the multi-label classification task. These metrics are not treated as final indicators but as the basis for further analysis. The evaluation stage is followed by a behavior-oriented analysis consisting of metric divergence examination, class-level comparison, and precision–recall pattern interpretation. These analyses are used to construct regime-based categorizations and provide a structured interpretation of model behavior under imbalanced multi-label conditions.

2.5. Regime Categorization Criteria

Classes are categorized into three reliability regimes based on their per-class F1 scores, applied consistently across both models:

1) High-confidence: F1 ≥ 0.55. Classes in this regime demonstrate stable and reliable classification performance with sufficient predictive consistency.
2) Unstable: 0.25 ≤ F1 < 0.55. Classes in this regime show moderate but variable performance that may fluctuate depending on threshold settings and data characteristics.
3) Non-functional: F1 < 0.25. Classes in this regime demonstrate near-zero predictive reliability, where the model effectively fails to classify the class correctly in a consistent manner.

The threshold values used for regime categorization (F1 = 0.25 and F1 = 0.55) were selected empirically based on the observed distribution of per-class F1 scores. Their purpose is to provide an interpretable separation between classes with high, moderate, and very low predictive reliability. These thresholds are not intended as universal

performance standards, but rather as an analytical framework to facilitate structured interpretation of class-level behavior under imbalanced multi-label conditions.

3. RESULTS AND DISCUSSION

3.1. Metric Divergence as a Behavioral Signature

The aggregate evaluation results are summarized in Table 4. In addition to Macro F1 and Micro F1, the table reports Macro Precision, Macro Recall, Micro Precision, Micro Recall, and Hamming Loss to provide a more complete view of model behavior under imbalanced multi-label conditions.

Table 4. Aggregate performance comparison of IndoBERT and DistilBERT

Model	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1	Difference	Hamming Loss
DistilBERT	0.4427	0.5411	0.4414	0.2097	0.6364	0.3155	+0.126	0.1150
IndoBERT	0.4900	0.4780	0.4659	0.5090	0.6240	0.5607	-0.095	0.0407

As presented in Table 4, the two models exhibit substantially different evaluation profiles. IndoBERT achieves higher Macro Precision (0.4900), Micro Precision (0.5090), Micro F1 (0.5607), and lower Hamming Loss (0.0407). In contrast, DistilBERT attains higher Macro Recall (0.5411) and Micro Recall (0.6364), indicating broader label activation across emotion categories. Notably, DistilBERT's Micro Precision of 0.2097 against a Micro Recall of 0.6364 confirms that approximately three out of every four positive label predictions made by DistilBERT are false positives, reflecting substantial over-prediction across the label space.

The difference between precision and recall provides an initial indication of prediction behavior. DistilBERT exhibits relatively high recall but substantially lower precision, particularly at the micro level, where Micro Precision reaches only 0.2097 despite a Micro Recall of 0.6364. This pattern suggests that the model tends to predict a larger number of labels, resulting in increased false positive assignments. IndoBERT shows a more balanced precision–recall profile, maintaining considerably higher precision while preserving competitive recall.

These differences are reflected in the relationship between Macro F1 and Micro F1. IndoBERT achieves a higher Micro F1 (0.5607) than Macro F1 (0.4659), indicating that its predictive performance is concentrated on high-frequency classes that dominate the dataset. DistilBERT exhibits the opposite pattern, where Macro F1 (0.4414) exceeds Micro F1 (0.3155), suggesting broader coverage across classes but lower instance-level reliability. The opposing directions of this divergence are visualized in Figure 2.

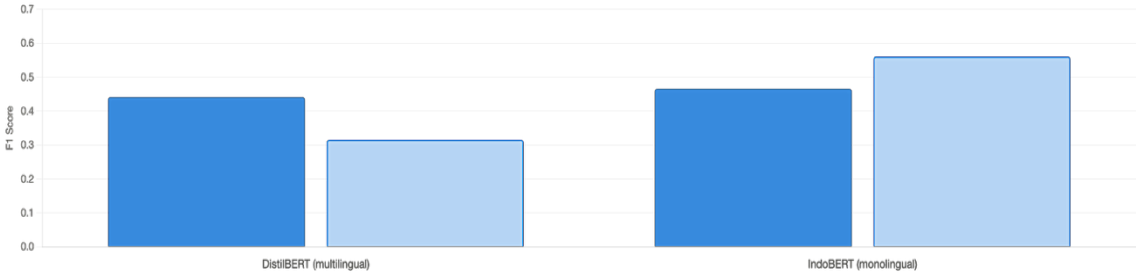

Figure 2. Macro F1 vs Micro F1 comparison for DistilBERT and IndoBERT

As shown in Figure 2, the direction of the Macro–Micro F1 gap differs between the two models. DistilBERT produces a positive gap of +0.126, whereas IndoBERT produces a negative gap of −0.095. This contrast suggests that the models allocate predictive capacity differently across the label space. DistilBERT distributes predictions more broadly across classes, while IndoBERT concentrates performance on labels with stronger representation in the dataset.

The Hamming Loss values further support this interpretation. IndoBERT achieves substantially lower Hamming Loss than DistilBERT, indicating fewer incorrect label assignments per instance. Taken together, the precision–recall profile, Macro–Micro F1 divergence, and Hamming Loss values provide initial evidence that the two models follow different prediction strategies, which are examined in greater detail through class-level analysis in the following sections.

3.2. Frequency – Performance Relationship

To further examine how metric divergence relates to data characteristics, per-class F1 scores are analyzed in relation to the number of test instances. For classes with more than approximately 100 to 150 instances, both models achieve relatively stable and moderate to high F1 scores. As the number of instances decreases, performance declines

more sharply. Below approximately 80 instances, most classes exhibit consistently low F1 scores across both models. This pattern suggests that sufficient data availability plays an important role in supporting stable classification performance. However, the relationship is not fully deterministic. Some classes deviate from the general pattern: fear achieves relatively high F1 despite limited data, possibly due to clearer lexical signals in the Indonesian translation, while realization shows lower performance despite having more instances, possibly due to semantic overlap with related categories such as surprise and understanding. This pattern is illustrated in Figure 3.

Figure 3. Scatter plot of IndoBERT and DistilBERT F1 scores against test instance count for 14 selected classes

Figure 3 presents the frequency–performance scatter plot for 14 selected emotion classes. These classes were chosen to represent the full frequency spectrum: the six highest-frequency classes, the six lowest-frequency classes with at least one test instance, and two mid-frequency classes selected for their behavioral deviation from the general trend. This selection enables visual clarity while preserving the interpretive range of the relationship. Complete per-class results for all 28 classes including optimized thresholds are provided in Table 5 (IndoBERT) and Table 6 (DistilBERT) below.

Table 5. Per-Class Performance of IndoBERT for All 28 Emotion Classes

Emotion	Support	Precision	Recall	F1	Regime
Neutral	1787	0.566	0.825	0.672	High-confidence
Admiration	504	0.673	0.601	0.635	High-confidence
Amusement	264	0.731	0.833	0.779	High-confidence

Emotion	Support	Precision	Recall	F1	Regime
Approval	351	0.324	0.444	0.375	Unstable
Caring	135	0.446	0.400	0.422	Unstable
Curiosity	284	0.419	0.813	0.553	High-confidence
Desire	83	0.455	0.422	0.438	Unstable
Disapproval	267	0.269	0.562	0.364	Unstable
Disappointment	156	0.617	0.455	0.524	Unstable
Disgust	123	0.711	0.260	0.381	Unstable
Embarrassment	37	0.550	0.297	0.386	Unstable
Excitement	103	0.548	0.330	0.412	Unstable
Fear	78	0.561	0.705	0.625	High-confidence
Gratitude	352	0.944	0.869	0.905	High-confidence
Grief	6	0.250	0.167	0.200	Non-functional
Joy	161	0.551	0.540	0.545	Unstable
Love	238	0.745	0.786	0.765	High-confidence
Nervousness	11	0.250	0.091	0.133	Non-functional
Optimism	186	0.510	0.548	0.528	Unstable
Pride	16	0.667	0.375	0.480	Unstable
Realization	145	0.256	0.138	0.179	Non-functional
Relief	11	0.250	0.091	0.133	Non-functional
Remorse	56	0.533	0.714	0.611	High-confidence
Sadness	156	0.617	0.455	0.524	Unstable
Surprise	141	0.508	0.454	0.479	Unstable
Anger	198	0.330	0.434	0.375	Unstable
Annoyance	320	0.323	0.375	0.347	Unstable
Confusion	153	0.437	0.340	0.382	Unstable

Table 6. Per-Class Performance of DistilBERT for All 28 Emotion Classes

Emotion	Support	Precision	Recall	F1	Regime
Neutral	1787	0.543	0.853	0.664	High-confidence
Admiration	504	0.603	0.675	0.637	High-confidence
Amusement	264	0.720	0.856	0.782	High-confidence
Approval	351	0.223	0.504	0.309	Unstable
Caring	135	0.444	0.296	0.356	Unstable
Curiosity	284	0.425	0.820	0.560	High-confidence
Desire	83	0.500	0.265	0.347	Unstable
Disapproval	267	0.281	0.420	0.336	Unstable

Emotion	Support	Precision	Recall	F1	Regime
Disappointment	156	0.549	0.506	0.527	Unstable
Disgust	123	0.522	0.390	0.447	Unstable
Embarrassment	37	0.500	0.405	0.448	Unstable
Excitement	103	0.286	0.388	0.329	Unstable
Fear	78	0.676	0.615	0.644	High-confidence
Gratitude	352	0.965	0.858	0.908	High-confidence
Grief	6	0.001	1.000	0.002	Non-functional
Joy	161	0.500	0.627	0.557	High-confidence
Love	238	0.732	0.769	0.750	High-confidence
Nervousness	11	0.002	1.000	0.004	Non-functional
Optimism	186	0.514	0.581	0.546	Unstable
Pride	16	0.750	0.375	0.500	Unstable
Realization	145	0.276	0.186	0.222	Non-functional
Relief	11	0.002	1.000	0.004	Non-functional
Remorse	56	0.465	0.821	0.594	High-confidence
Sadness	156	0.549	0.506	0.527	Unstable
Surprise	141	0.469	0.376	0.417	Unstable
Anger	198	0.337	0.338	0.338	Unstable
Annoyance	320	0.323	0.328	0.326	Unstable
Confusion	153	0.381	0.399	0.390	Unstable

3.3. Class-Level Comparison and The Limits of Aggregate Superiority

While aggregate metrics suggest that IndoBERT performs better overall, this advantage is not uniformly distributed across emotion categories. To examine how the two models differ at the class level, their per-class F1 scores are compared directly. Figure 4 presents the F1 gap between IndoBERT and DistilBERT across 14 selected emotion categories, where positive values (blue bars) indicate classes where IndoBERT leads and negative values (orange bars) indicate classes where DistilBERT achieves higher F1. Classes with near-zero gaps appear close to the center axis.

As shown in Figure 4, IndoBERT's advantage is concentrated in a distinct subset of classes. The largest positive gaps are observed for relief, nervousness, excitement, and caring, suggesting that IndoBERT is considerably more effective on these categories. Moderate positive gaps are also visible for disappointment and remorse. In contrast, DistilBERT

outperforms IndoBERT on disgust, embarrassment, realization, and fear, as indicated by the negative bars. For mid-frequency classes such as neutral, admiration, gratitude, and amusement, the gap is minimal, appearing as very short or dashed bars close to zero, indicating comparable performance. Classes with very low frequency tend to exhibit near-zero F1 for both models, making the gap between them less interpretively meaningful in this region. These results indicate that aggregate superiority does not imply consistent improvement across all classes. The two models exhibit complementary strengths depending on class frequency and label characteristics. To understand the underlying mechanism driving these differences, precision and recall are examined separately in the following section.

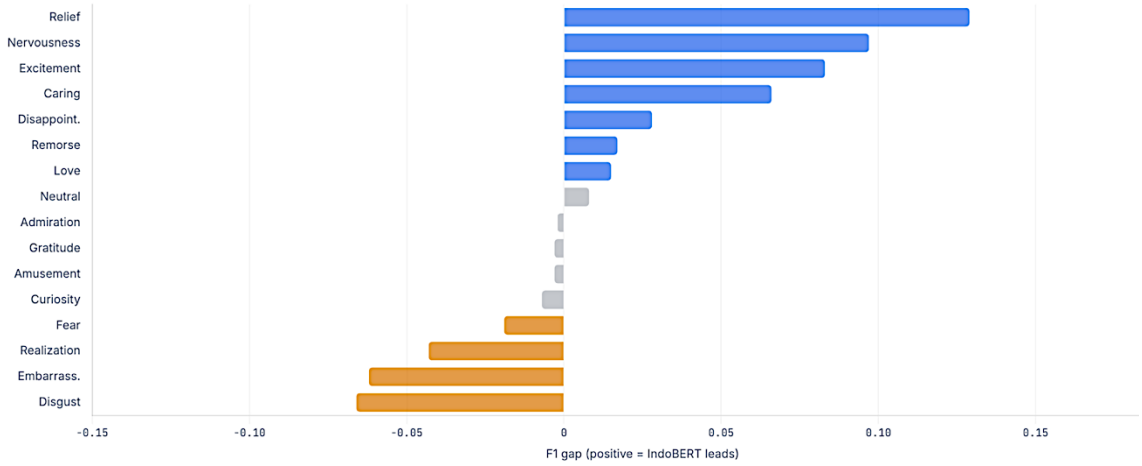

Figure 4. Per-class F1 difference (IndoBERT and DistilBERT) across emotion categories.

3.4. Precision and Recall Dynamics

To further understand the class-level differences observed in Figure 4, precision and recall are examined separately for each model. While F1-score combines both metrics into a single value, examining them independently reveals how each model makes prediction decisions, namely whether it tends to be conservative and selective with high precision and lower recall, or broad and inclusive with high recall and lower precision. Figures 5a and 5b present the per-class precision and recall values for IndoBERT and DistilBERT, respectively, across selected emotion categories.

Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Figure 5a. IndoBERT Precision and Recall

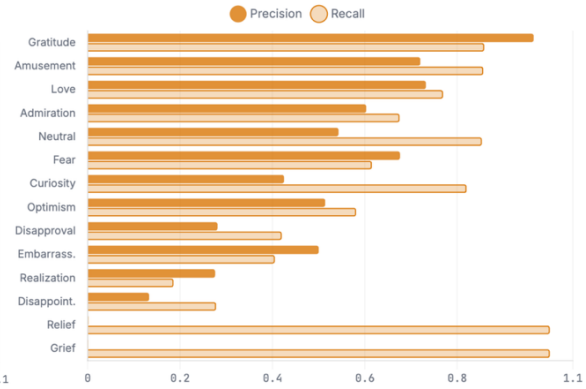

Figure 5b. DistilBERT Precision and Recall

As shown in Figures 5a and 5b, the two models exhibit distinctly different prediction patterns. IndoBERT consistently maintains higher precision across most classes, with gratitude, amusement, love, and admiration all achieving precision values above 0.6. However, recall for IndoBERT drops notably for several classes such as curiosity, neutral, and fear, indicating that the model tends to under-predict these categories and miss true positive instances in order to reduce false positives. DistilBERT shows the opposite tendency. Recall values are generally higher across classes including curiosity, neutral, fear, and grief, reflecting broader label activation. However, precision for DistilBERT is comparatively lower in many categories such as embarrassment, disapproval, and realization, suggesting a higher number of false positive predictions. For low-frequency classes such as grief and relief, both models achieve near-zero precision, confirming that data scarcity severely limits prediction reliability regardless of architecture.

This precision–recall trade-off directly explains the Macro–Micro F1 divergence observed in Section 3.1. IndoBERT's selective behavior, characterized by high precision and conservative recall, produces higher Micro F1 by concentrating accurate predictions on high-frequency classes that dominate instance-level aggregation. In contrast, DistilBERT exhibits broader label activation with higher recall but lower precision, distributing prediction effort more evenly across classes. This behavior improves Macro F1 but reduces Micro F1 because false positive predictions accumulate across the label space. These contrasting tendencies indicate that the direction of Macro–Micro F1 divergence provides meaningful insight into each model's underlying prediction strategy.

3.5. Regime-Based Structure of Performance

The variation in class-level performance can be further understood by grouping classes into distinct reliability regimes. Based on the criteria defined in Section 2.5, classes are categorized into three groups: high-confidence, unstable, and non-functional.

Table 7. Regime Membership for IndoBERT and DistilBERT Across All 28 Emotion Classes

Regime	IndoBERT Classes	DistilBERT Classes
High-confidence (F1 ≥ 0.55)	neutral, admiration, amusement, curiosity, fear, gratitude, love, remorse	neutral, admiration, amusement, curiosity, fear, gratitude, joy, love, remorse
Unstable (0.25 ≤ F1 < 0.55)	approval, caring, desire, disapproval, disappointment, disgust, embarrassment, excitement, joy, optimism, pride, sadness, surprise, anger, annoyance, confusion	approval, caring, desire, disapproval, disappointment, disgust, embarrassment, excitement, optimism, pride, sadness, surprise, anger, annoyance, confusion
Non-functional (F1 < 0.25)	grief, nervousness, realization, relief	grief, nervousness, realization, relief

As shown in Table 7, the non-functional regime contains the same four classes (grief, nervousness, realization, relief) for both models, confirming that these classes are primarily constrained by data characteristics rather than architectural differences. The high-confidence regime is slightly larger for IndoBERT, while DistilBERT reassigns the approval class to the unstable regime due to its broader but less precise activation pattern. The similarity of regime structure across models underscores that class-level reliability is predominantly driven by data availability and label clarity rather than model architecture. The average F1 scores within each regime for both models are presented in Figure 6. This visualization complements Table 4 by showing not only which classes belong to each regime, but also the magnitude of performance differences between regimes.

As shown in Figure 6, a clear and consistent separation exists between the three regimes across both models. High-confidence classes achieve average F1 scores of approximately 0.69 for IndoBERT and 0.67 for DistilBERT, confirming stable and reliable prediction in these categories. Unstable classes show intermediate average F1 values of around 0.41 for IndoBERT and 0.40 for DistilBERT, reflecting variable but partially functional

performance. Non-functional classes remain well below 0.25, with average F1 scores of approximately 0.18 for IndoBERT and 0.13 for DistilBERT, indicating near complete failure to predict these categories reliably. The near-identical regime structure across both models reinforces the finding that these performance tiers are primarily shaped by data characteristics particularly class frequency and label clarity rather than by the choice of model architecture.

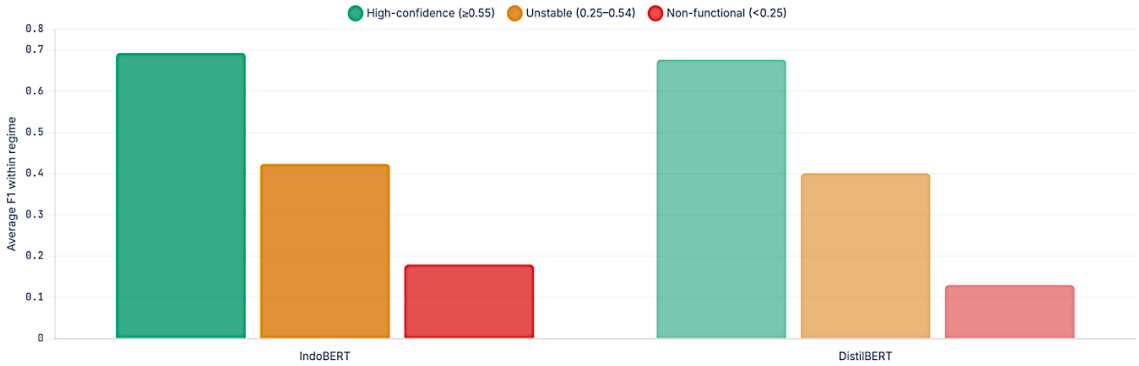

Figure 6. Average F1-scores across performance regimes for IndoBERT and DistilBERT

3.6. Distribution of Per-Class Performance

A broader perspective on class-level behavior can be obtained by examining the overall distribution of per-class F1 scores across all 28 emotion categories. While the regime-based analysis in Section 3.5 groups classes into discrete categories, the histograms in Figures 7 and 8 reveal the underlying shape of the performance distribution for each model, showing how many classes fall within each F1 score interval.

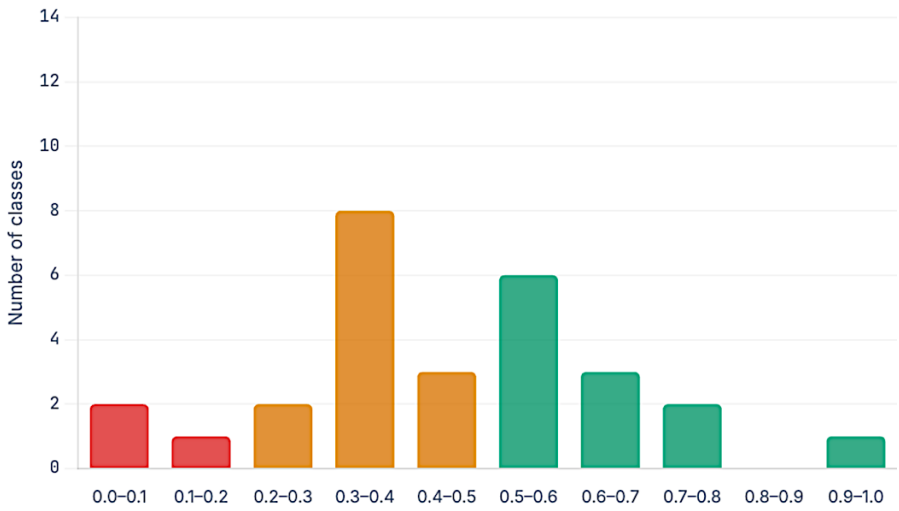

Figure 7. Histogram of IndoBERT Per-Class F1 Across All 28 Classes

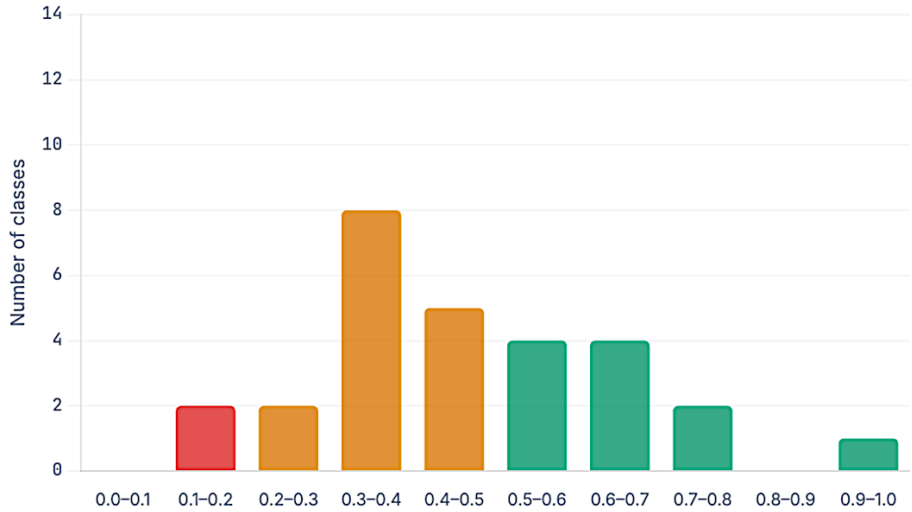

Figure 8. DistilBERT Per-Class F1 Histogram Per-Class F1 Across All 28 Classes

As shown in Figures 7 and 8, the distribution of F1 scores is notably non-uniform for both models, confirming that performance does not form a continuous spectrum but instead clusters into distinct bands. For IndoBERT (Figure 7), the largest concentration of classes falls in the 0.3–0.4 interval with 8 classes, followed by a secondary peak in the 0.5–0.6 range with 6 classes, and a smaller cluster at 0.6–0.7 with 3 classes. A few classes achieve high F1 scores above 0.7, while only 3 classes fall below 0.2, corresponding to the non-functional regime. For DistilBERT (Figure 8), the distribution shows a similar dominant peak at 0.3–0.4 with 8 classes, but the spread into higher F1 intervals is more gradual: 5 classes in 0.4–0.5, 4 classes each in 0.5–0.6 and 0.6–0.7, and 2 classes in 0.7–0.8. Notably, DistilBERT has no classes in the 0.0–0.1 range, whereas IndoBERT has 2, suggesting that DistilBERT's broader activation prevents complete failure even for the most difficult classes, though at the cost of lower precision.

These distributional patterns reinforce the regime-based interpretation: performance naturally clusters into high, mid, and low ranges rather than distributing uniformly. IndoBERT's distribution is more bimodal, concentrated at the lower-mid range and the upper-mid range reflecting its selective strategy of performing well on a subset of classes while largely ignoring others. DistilBERT's distribution is more spread across mid-level intervals, consistent with its broader but less precise label activation. Taken together, these histograms confirm that aggregate metrics summarize performance across a highly heterogeneous distribution of class-level outcomes, further supporting the need for class-level analysis in imbalanced multi-label evaluation.

3.7. Discussion

The results of this study indicate that differences between Macro F1 and Micro F1 reflect how models respond to imbalanced data distributions rather than simply representing variations in performance scale. The opposing divergence patterns observed in IndoBERT and DistilBERT confirm that each model allocates predictive capacity differently across the label space, consistent with the behavioral interpretation framework proposed in this study.

The opposing divergence patterns can be attributed to several interacting factors. For IndoBERT, the higher Micro F1 relative to Macro F1 is likely driven by its monolingual pre-training on Indonesian corpora: having been trained exclusively on Indonesian text, IndoBERT develops strong representations for frequent, semantically clear Indonesian emotion expressions such as neutral, gratitude, and admiration. However, this advantage does not transfer effectively to low-frequency classes, where limited training examples prevent the model from learning reliable class boundaries. For DistilBERT, the higher Macro F1 relative to Micro F1 may reflect the effect of multilingual pre-training across 104 languages: exposure to emotion-bearing text in multiple languages may support more generalizable representations that activate across a broader set of emotion categories, even with fewer Indonesian-specific examples. However, this broader activation comes at the cost of precision, resulting in more false positives and a lower Micro F1. Additionally, differences in class-specific threshold optimization may have contributed to the observed recall–precision trade-off between the two models. IndoBERT required substantially lower thresholds for many minority classes (range: 0.10–0.75), whereas DistilBERT produced a narrower and generally higher threshold distribution (range: 0.50–0.625). This pattern indicates that the two models required different decision boundaries to optimize class-level performance, particularly for low-frequency emotions. However, because threshold values and pre-training characteristics vary simultaneously across the two models, their individual effects on recall and precision cannot be isolated in the current experimental design.

These results can be contextualized with prior work on GoEmotions. Demszky [11] reported Macro F1 scores of 0.46–0.50 using transformer-based models on the original English dataset, a range that closely aligns with the values observed here (0.44 for DistilBERT, 0.47 for IndoBERT). This suggests that translating GoEmotions into Indonesian

does not fundamentally alter the difficulty of the classification task at the aggregate level. However, the per-class analysis reveals patterns that aggregate comparisons cannot capture. Prior Indonesian NLP studies have demonstrated the effectiveness of IndoBERT across multiple benchmark tasks [13], [27].

In practical deployment, DistilBERT's broader recall behavior implies a higher rate of false positive label assignments. For applications where over-prediction is costly such as content moderation or clinical emotion monitoring where false positive flags may trigger unnecessary review processes, IndoBERT's more selective behavior may be preferable. Conversely, for applications where coverage is prioritized over precision such as exploratory emotion analysis or user experience research, DistilBERT's broader activation may be acceptable. This distinction underlines the importance of selecting models and evaluation criteria based on downstream task requirements rather than aggregate metric scores alone. The regime-based framework introduced in this study provides a practical tool for making such distinctions explicit, enabling more informed model selection for downstream Indonesian emotion classification tasks.

The influence of translation quality on the results also warrants consideration. For low-frequency and semantically subtle emotions such as grief, nervousness, and pride, machine translation may introduce noise or alter the semantic boundaries between adjacent categories. This may partly explain why these classes consistently fall into the non-functional regime across both models: the combined effect of limited data and potentially imprecise translation makes reliable learning difficult. Future work using native Indonesian emotion data would help isolate the contribution of translation artifacts from data availability effects.

Future work on improving minority-class reliability may explore several directions. Class-balanced loss functions or focal loss can reweight the training objective to prioritize minority classes during optimization. Data augmentation strategies such as back-translation or synonym replacement may increase the effective number of minority-class training examples. Label grouping consolidating semantically overlapping low-frequency classes into broader categories may reduce sparsity and improve learnability. Alternatively, hierarchical classification frameworks that first predict coarse emotion

categories before fine-grained labels may mitigate the cold-start problem for rare classes.

This study has several important limitations. First, the dataset is based on machine translation of English Reddit comments into Indonesian, which may not accurately preserve culturally specific emotional expressions, particularly for fine-grained and low-frequency categories. Second, emotional label assignments from the original English annotation are assumed to remain valid after translation without re-validation, which may not hold for all 28 categories. Third, the experiment was conducted in a single run with a fixed random seed; cross-validation was not applied, meaning results may be sensitive to the specific data split. Fourth, threshold optimization was performed on the validation set, introducing a risk of threshold overfitting that may not generalize equally to unseen data. Fifth, the comparison is limited to two transformer architectures, which constrains the generalizability of the behavioral patterns observed. Finally, the statistical significance of the observed performance differences between models was not formally tested, as the single-run design does not support confidence interval estimation.

4. CONCLUSION

This study proposes a behavior-oriented interpretation framework for understanding Macro–Micro F1 divergence in imbalanced multi-label Indonesian emotion classification. Using a machine-translated Indonesian GoEmotions dataset containing 28 emotion classes, the comparative analysis of IndoBERT and DistilBERT demonstrates that metric divergence provides meaningful diagnostic signals regarding how predictive performance is distributed across the label space. Although both models achieve comparable aggregate performance, they exhibit fundamentally different prediction behaviors: IndoBERT concentrates performance on high-frequency classes, whereas DistilBERT distributes predictions more broadly with lower precision. To further characterize these differences, a regime-based class reliability categorization is introduced, revealing that model behavior is distributed across high-confidence, unstable, and non-functional performance regimes rather than representing a uniform capability across all classes. The similar regime structures observed across both models suggest that class frequency and label characteristics play a stronger role than model architecture in shaping class-level performance patterns. These findings indicate that

aggregate metrics alone are insufficient for evaluating imbalanced multi-label systems and should be complemented by class-level behavioral analysis. Nevertheless, the results should be interpreted within the context of a machine-translated Indonesian GoEmotions dataset that was not manually revalidated, a two-model comparison, and a single experimental setting. Future research may extend this framework using native Indonesian emotion datasets, broader model comparisons, and alternative imbalance-mitigation strategies to further examine the generalizability of the observed behavioral patterns.

REFERENCES

[1] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[2] S. Ossenov, "Developing a Dataset-Adaptive, Normalized Metric for Machine Learning Model Assessment: Integrating Size, Complexity, and Class Imbalance," arXiv preprint arXiv: 2412.07244, 2024. Accessed: May 27, 2026. [Online]. Available: https://arxiv.org/abs/2412.07244

[3] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences*, vol. 14, no. 21, p. 9863, Oct. 2024, doi: 10.3390/app14219863.

[4] S. Roohi, R. Skarbez, and H. D. Nguyen, "Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework," *Natural Language Processing*, vol. 31, no. 5, pp. 1163–1186, Sep. 2025, doi: 10.1017/nlp.2024.48.

[5] D. Harbecke, Y. Chen, L. Hennig, and C. Alt, "Why only Micro-F1? Class Weighting of Measures for Relation Classification," in *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, T. Shavrina, V. Mikhailov, V. Malykh, E. Artemova, O. Serikov, and V. Protasov, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 32–41. doi: 10.18653/v1/2022.nlppower-1.4.

[6] Y. Xia, Q. Zhao, Y. Long, G. Xu, and J. Wang, "SensoryT5: Infusing Sensorimotor Norms into T5 for Enhanced Fine-grained Emotion Classification," in Proc. Workshop on Cognitive Aspects of the Lexicon (CogALex), Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 144–152, doi: 10.18653/v1/2024.cogalex-1.19.

[7] B. Pithava, A. Magar, and S. Bharti, "Unveiling Sentiment Dynamics: Emotion Detection in Social Media," in *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/ICEC59683.2024.10837523.

[8] Z. Su, H. Lyu, Y. Niu, and Y. Liu, "Based on Data Balancing and Model Improvement for Multi-Label Sentiment Classification Performance Enhancement," arXiv preprint arXiv: 2511.14073, 2025, Accessed: May 27, 2026. [Online]. Available: https://arxiv.org/abs/2511.14073

[9] R. Chauhan, A. Gusain, P. Kumar, C. Bhatt, and I. Uniyal, "Fine Grained Sentiment Analysis using Machine Learning and Deep Learning," in *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*, IEEE, Sep. 2023, pp. 423–427. doi: 10.1109/ICSEIET58677.2023.10303481.

[10] A. Sharma, A. Avasthi, V. L. Vangipuram, P. G., S. V., and T. C. Manjunath, "Exploring Emotion Psychology in AI: Common Perspectives and Their Application in Research and Development to Enhance Empathetic Responses in Artificial Intelligence Systems," in *2025 7th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, Sep. 2025, pp. 1–6. doi: 10.1109/ISCON65210.2025.11341720.

[11] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4040–4054, doi: 10.18653/v1/2020.acl-main.372.

[12] L. Piras, L. Boratto, and G. Ramos, "Evaluating the Prediction Bias Induced by Label Imbalance in Multi-label Classification," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, New York, NY, USA: ACM, Oct. 2021, pp. 3368–3372. doi: 10.1145/3459637.3482100.

[13] 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*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[14] M. R. Syazali and E. Yulianti, "Classification of Economic Activities in Indonesia Using IndoBERT Language Model," *Jurnal Ilmu Komputer dan Informasi*, vol. 18, no. 2, pp. 155–165, Jun. 2025, doi: 10.21609/jiki.v18i2.1446.

[15] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, Dec. 2025, doi: 10.1007/s13278-025-01439-6.

[16] W. Wongso, D. S. Setiawan, S. Limcorn, and A. Joyoadikusumo, "NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural," in Proc. Second Workshop in South East Asian Language Processing (SEALP), Online: Association for Computational Linguistics, Jan. 2025, pp. 10–26, doi: 10.18653/v1/2025.sealp-1.2.

[17] W. Christian, D. Adamlu, A. Yu, and D. Suhartono, "Leveraging IndoBERT and DistilBERT for Indonesian emotion classification in e-commerce reviews," *Procedia Comput. Sci.*, vol. 269, pp. 321–330, 2025, doi: 10.1016/j.procs.2025.08.284.

[18] E. I. Setiawan, L. Kristianto, A. T. Hermawan, J. Santoso, K. Fujisawa, and M. H. Purnomo, "Social Media Emotion Analysis in Indonesian Using Fine-Tuning BERT Model," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, IEEE, Apr. 2021, pp. 334–337. doi: 10.1109/EIConCIT50028.2021.9431885.

[19] S. Goldfarb-Tarrant, B. Ross, and A. Lopez, "Cross-lingual Transfer Can Worsen Bias in Sentiment Analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5691–5704. doi: 10.18653/v1/2023.emnlp-main.346.

[20] J. Li *et al.*, "A Two-Stage Framework for Ambiguous Classification in Software Engineering," in *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, IEEE, Oct. 2023, pp. 275–286. doi: 10.1109/ISSRE59848.2023.00070.

[21] A. Amalia, M. S. Lydia, P. I. Nainggolan, Nurrahmadayeni, S. Br Siagian, and D. S. Br Ginting, "Multi-Label Emotion Classification for Indonesian Text using IndoBERT Fine-Tuning," in *2025 9th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, IEEE, Nov. 2025, pp. 293–299. doi: 10.1109/ELTICOM67568.2025.11336043.

[22] R. Kumar, R. K. Ayyasamy, and A. K. Jebna, "Long-Tail Emotion Detection: Few-Shot Learning for Rare Pandemic Emotions via Prototype Networks," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 55, no. 1, pp. 236–244, Aug. 2025, doi: 10.37934/araset.55.1.236244.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[23] N. V. S. J. Jami et al., “Stratify or Die: Rethinking Data Splits in Image Segmentation,” arXiv preprint arXiv: 2509.21056, 2025. Accessed: May 27, 2026. [Online]. Available: https://arxiv.org/abs/2509.21056

[24] T. T. Inan, M. Liu, and A. Shehu, "F-Measure Optimization for Multi-class, Imbalanced Emotion Classification Tasks," in Artificial Neural Networks and Machine Learning – ICANN 2022, Lecture Notes in Computer Science, vol. 13529, Springer, 2022, pp. 158–170, doi: 10.1007/978-3-031-15919-0_14.

[25] S. Simhadri, M. Ponnam, R. Rajitha, and R. Balamurugan, "Enhanced Multi-Class Model Evaluation: Analyzing BERT, GPT-2, and LLaMA with Precision, Recall, and F1-Score Metrics," in Proc. 4th Int. Conf. Innovative Mechanisms for Industry Applications (ICIMIA), IEEE, 2025, pp. 984–989, doi: 10.1109/ICIMIA67127.2025.11200914.

[26] R. Vinston Raja et al., “Metrics and Techniques for Evaluating Machine Learning Models and Optimization Algorithms,” in AI Model Design and Data Management for Disease Prediction, A. Muniasamy, Ed., IGI Global Scientific Publishing, 2025, pp. 193–222, doi: 10.4018/979-8-3373-5137-7.ch007.

[27] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in Proc. Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th Int. Joint Conf. Natural Language Processing (AACL-IJCNLP), 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.