Addressing Extreme Class Imbalance in Multilingual Complaint Classification Using XLM-RoBERTa

Muhammad Ariyanto [1]*, Farrikh Alzami [2]*, Ramadhan Rakhmat Sani [3]*, Indra Gamayanto [4]*, Muhammad Naufal [5], Sri Winarno [6]**, Iswahyudi [7]*****
* Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
** Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
*** Dinas Komunikasi dan Informatika Provinsi Jawa Tengah
112202206880@mhs.dinus.ac.id [1], alzami@dsn.dinus.ac.id [2], ramadhan_rs@dsn.dinus.ac.id [3], indra.gamayanto@dsn.dinus.ac.id [4], m.naufal@dsn.dinus.ac.id [5], sri.winarno@dsn.dinus.ac.id [6], iswahyudi@jatengprov.go.id [7]

Article Info

Article history:

Received 2025-10-30
Revised 2025-11-25
Accepted 2025-12-22

Keyword:

Class Imbalance,
XLM-RoBERTa,
E-Government,
Indonesian NLP,
Text Classification.

ABSTRACT

Government complaint management systems often suffer from extreme class imbalance, where a few public service categories accumulate most reports while many others remain under-represented. This research examines whether simple class weighting can improve fairness in multilingual transformer models for automatic routing of Indonesian citizen complaints on the LaporGub Central Java e-governance platform. The dataset comprises 53,877 Indonesian-language complaints spanning 18 service categories with an imbalance ratio of about 227:1 between the largest and smallest classes. After cleaning and deduplication, we stratify the data into training, validation, and test sets. We compare three approaches: (i) a linear support vector machine (SVM) with term frequency inverse document frequency (TF-IDF) unigram and bigram and class-balanced weights, (ii) a cross-lingual RoBERTa (XLM-RoBERTa-base) model without class weighting, and (iii) an XLM-RoBERTa-base model with a class-weighted cross-entropy loss. Fairness is operationalised as equal importance for categories and quantified primarily using the macro-averaged F1-score (Macro-F1), complemented by per-class F1, weighted F1, and accuracy. The unweighted XLM-RoBERTa model outperforms the SVM baseline in Macro-F1 (0.610 vs 0.561). The class-weighted variant attains similar Macro-F1 (0.608) while redistributing performance towards minority categories. Analysis shows that class weighting is most beneficial for categories with a few hundred to several thousand samples, whereas extremely rare categories with fewer than 200 complaints remain difficult for all models and require additional data-centric interventions. These findings demonstrate that multilingual transformer architectures combined with simple class weighting can provide a more balanced backbone for automated complaint routing in Indonesian e-government, particularly for low- and medium-frequency service categories.

I. INTRODUCTION

Digital transformation has redefined public service delivery, particularly in the domain of citizen complaint management. Modern *e-governance* systems allow citizens to submit, track, and evaluate complaints transparently through web-based platforms that enhance efficiency, accountability, and public trust [1]. Compared to paper-based or in-person processes that are slow and error-prone, digital complaint systems ensure traceability and faster response times [2]. Such systems have become integral to participatory governance, especially in developing regions like Indonesia, where technology-based reporting supports open communication and responsiveness in local administrations [3].

However, despite the adoption of digital complaint portals, most classification processes remain manually handled. Human officers must read and assign categories to thousands of submissions an approach that is labor-intensive, inconsistent, and unsuitable for large-scale citizen participation [2]. This growing volume of textual data highlights the need for automated language understanding systems. Early solutions using traditional models such as Support Vector Machines (SVMs) and TF-IDF achieved reasonable accuracy but required manual feature design and lacked contextual comprehension [4]. These models failed to capture semantic relations or informal expressions common in citizen narratives, limiting their performance in real-world complaint datasets [5].

Recent advances in deep learning have revolutionized text classification by enabling models to learn semantic representations directly from raw text. Transformer-based architectures such as mBERT and XLM-RoBERTa have demonstrated superior cross-lingual generalization and contextual understanding [6]. XLM-RoBERTa, for instance, surpasses mBERT by more than 14% accuracy on multilingual benchmarks and remains robust for low-resource languages [6]. Further innovations such as prompt-based fine-tuning have proven that multilingual transformers can achieve language-independent accuracy even with limited labeled data [7], [8]. These models have become the global standard for multilingual NLP tasks, showing scalability and effectiveness across diverse languages and domains.

Parallel developments in the Indonesian NLP ecosystem further validate the power of transformer architectures. Studies on sentiment and emotion classification reveal that hybrid deep-learning strategies combining BERT and DistilBERT with sequential models outperform conventional approaches on Indonesian datasets [9]. Fine-tuned IndoBERT models for social media sentiment analysis achieve notable accuracy and F1-score improvements over baselines [10], while IndoBERT also performs strongly on emotion classification tasks in e-commerce reviews [11]. These advancements are underpinned by the establishment of major language resources such as IndoLEM [12], IndoNLU [13] and NusaBERT [14], which provide standardized benchmarks and pre-trained models for Indonesian NLP. Collectively, these resources enable reproducible evaluation and foster broader adoption of transformer models in low-resource Southeast Asian languages.

A key obstacle in applying NLP to public complaint systems lies in *extreme class imbalance*. Government datasets are often heavily skewed, where majority categories such as infrastructure or health dominate, while minority classes covering sensitive issues like corruption or gender-based violence remain underrepresented [3]. This imbalance leads to biased learning, with models favoring frequent categories and neglecting rare but critical topics [15]. Classical balancing techniques such as oversampling, undersampling, and synthetic data generation (e.g., SMOTE, ADASYN) have been used to address this issue [16], yet they often distort linguistic patterns or cause overfitting. Alternatively, algorithmic adjustments such as cost-sensitive learning, class weighting, and focal loss modify training objectives to emphasize minority classes, improving fairness and recall [17], [18], [19]. However, these methods show diminishing returns when sample sizes are extremely small, suggesting that imbalance beyond ratios of 200:1 remains a persistent challenge even for advanced models [19].

Within this context, multilingual transformer models offer a promising foundation for fair and scalable classification of Indonesian complaint data. Nevertheless, there is a lack of systematic empirical evidence comparing traditional baselines and transformer-based models under real-world extreme imbalance conditions in the Indonesian e-government domain. Furthermore, the threshold at which class weighting ceases to be effective particularly for ultra-minority categories has not been formally investigated. This research addresses these gaps by conducting an empirical evaluation of multilingual transformers against classical baselines for complaint classification in Indonesian public-service data. Specifically, it examines the impact of class weighting on minority-class performance, identifies the empirical sample-size boundary where weighting remains effective, and analyzes recurrent error patterns arising from semantic overlap and contextual ambiguity. Through this integration of multilingual representation learning and imbalance-aware modeling, the research contributes practical insights and reproducible benchmarks for developing fair, adaptive, and deployment-ready NLP systems in Indonesia's digital governance landscape.

II. METHOD

A. Research Stages

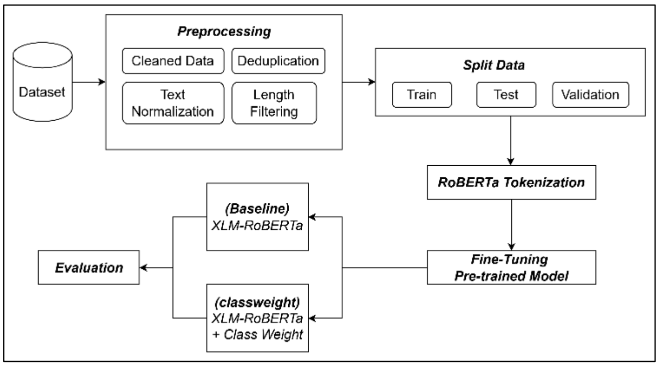

Figure 1. Research stages illustrating the sequential process

This research follows a structured empirical workflow designed to evaluate the effectiveness of class weighting in addressing extreme class imbalance in multilingual complaint classification. This process consists of several sequential stages: data collection, data preprocessing, stratified splitting, model training, evaluation, and analysis of results, as illustrated in Figure 1. Each stage is implemented in a controlled environment to ensure reproducibility and methodological transparency [19], [20].

B. Dataset

This research employed complaint reports from LaporGub Central Java, an official e-governance platform that enables citizens to report public service issues. The dataset is private and was accessed legally under a data use agreement, with all personally identifiable information removed prior to analysis. After sequential cleaning and deduplication, the final corpus comprised 53,877 labeled complaints distributed across 18 categories, with the largest class (Infrastruktur) containing

17,935 samples and the smallest (Saberpungli) 79 samples, resulting in an imbalance ratio of approximately 227:1.

To better characterise this long-tailed distribution, we group the 18 categories into four frequency tiers based on their counts in the full dataset. Tier 1 contains very frequent categories with more than 5,000 complaints (three classes: Infrastruktur, Sosial Masyarakat, and Pendidikan), Tier 2 includes ten medium-frequency categories with 1,000 to 5,000 complaints, and Tier 3 comprises four minority categories with 200 to 1,000 complaints. Tier 4 consists of a single extremely rare category, Saberpungli, with only 79 complaints. The lower boundary of approximately 200 samples reflects both a natural break in the empirical frequency distribution and an empirical point at which all models begin to exhibit unstable F1-scores and near-zero recall. We therefore use 200 examples as a practical threshold between "rare but learnable" and "extremely rare" classes in this dataset, rather than as a universal theoretical cutoff. [3], [15].

C. Preprocessing

Text preprocessing standardized the complaint data while preserving semantic integrity. Each record was converted to lowercase, and regular expressions were used to remove HTML tags, URLs, and email addresses, followed by whitespace normalization. No stemming, lemmatization, or stopword removal was applied to maintain grammatical and contextual cues essential to Indonesian complaint narratives [5] [9]. Duplicate text–category pairs were removed to prevent data leakage, and all preprocessing operations were implemented in Python using pandas and built-in regular expression functions.

A descriptive analysis of text length was conducted to determine an appropriate truncation limit for transformer inputs. The average complaint length was approximately 62 words, with a median of 45 and a standard deviation of 62, as illustrated in Figure 2. Based on this distribution, the maximum sequence length was set to 128 tokens, which captured the majority of complaint texts while minimizing unnecessary padding during model training [6], [12].

D. Stratified Splitting

To ensure balanced representation of all categories, the dataset was divided using stratified sampling into 70 % training (37 713 samples), 15 % validation (8 082 samples), and 15 % testing (8 082 samples). Stratification preserved proportional representation across classes, ensuring that minority categories such as Saberpungli and Pariwisata dan Budaya were included in all subsets, maintaining statistical validity and reducing sampling bias [16], [21].

E. Experimental Design

Three experimental configurations were established to assess the influence of class weighting on classification performance:

1. *Baseline 1 SVM + TF-IDF*
   A linear SVM trained on TF-IDF bigrams (10 000 features) with class weights inversely proportional to class frequency [4].
2. *Baseline 2 XLM-RoBERTa (No Class Weighting)*
   A multilingual transformer fine-tuned on the dataset without imbalance treatment [6].
3. *Proposed XLM-RoBERTa + Class Weighting*
   Identical to Baseline 2 but employing a class-weighted loss function to mitigate extreme imbalance.

The class weight for each class c was computed as:

$$\omega c = \frac{C\,x\,N}{Nc} \qquad (1)$$

where ωc denotes the class weight, C is the number of classes, N is the total number of samples, and Nc is the number of samples in class c. This weighting scheme increases the penalty for under-represented categories, encouraging balanced learning across all classes. All experiments shared identical data splits, preprocessing, and evaluation protocols to ensure comparability.

F. Model Configuration

All experiments were conducted on the Kaggle platform equipped with dual NVIDIA Tesla T4 GPUs. The XLM-RoBERTa-base model was fine-tuned using the Hugging Face Transformers library with the AdamW optimizer and a linear-warmup scheduler (10 % of total steps) [6], [20], [22].

TABLE I
MODEL CONFIGURATION AND HYPERPARAMETERS

Model	Representation	Parameter	Optimizer	Loss Function
SVM + TF-IDF	TF-IDF (1, 2) -gram, 10000 features	Linear SVC, max_iter = 2000, class_weight = balanced	-	Hinge Loss
XLM-RoBERTa	Transformer embeddings	LR = 2e-5, Batch = 16, Epochs = 10, MaxLen = 128, Patience = 3	AdamW + Linear Warmup (10 %)	CrossEntropyLoss
XLM-RoBERTa + Class Weighting		Same + Class Weights Eq.(1)		Weighted CrossEntropyLoss

Early stopping was triggered after three epochs without validation Macro-F1 improvement. Table I summarizes the main configuration parameters for all models. All implementations employed PyTorch, Transformers, and scikit-learn, ensuring consistent experimental environments and reproducibility across baselines [14], [20].

G. Evaluation Metrics

Model performance was primarily evaluated using Macro-F1, which computes the unweighted mean of per-class F1-scores and is robust to imbalance effects [15] . Secondary metrics included per-class Precision, Recall, Weighted-F1, and Accuracy, the latter reported with caution due to its sensitivity to dominant categories. Statistical robustness was assessed through bootstrap resampling (1,000 iterations) to estimate 95% confidence intervals for Macro-F1 scores.

In this research, we operationalise fairness as giving each complaint category equal importance via Macro-F1 and monitoring how performance changes for minority versus majority classes through per-class and grouped analyses. To summarise the relationship between class frequency and performance, we also report mean F1 within the four frequency tiers defined in Section II-B. These tiers are purely descriptive and are derived from the empirical class distribution; they are used only for analysis and do not affect how the models are trained. [3], [17].

III. Result And Discussion

A. Dataset Characteristic

TABLE II
Class Distribution Across 18 Complaint Categories with Tier Grouping Based On Sample Size

Category	Count	Percentage	Tier
Infrastruktur	17935	33.29	Tier 1
Sosial Masyarakat	9074	16.84	Tier 1
Pendidikan	6210	11.53	Tier 1
Permades Dan Kependudukan	3450	6.4	Tier 2
Kesehatan	2242	4.16	Tier 2
Kategori Lain-Lain	2068	3.84	Tier 2
Administrator	1964	3.65	Tier 2
Forkominda	1888	3.5	Tier 2
Energi	1809	3.36	Tier 2
Ekonomi Dan Industri	1303	2.42	Tier 2
Lingkungan	1283	2.38	Tier 2
Keuangan Dan Aset	1276	2.37	Tier 2
Kepegawaian	1047	1.94	Tier 2
Pertanian	878	1.63	Tier 3
Bencana	743	1.38	Tier 3
Pembangunan Daerah	380	0.71	Tier 3
Pariwisata Dan Budaya	248	0.46	Tier 3
Saberpungli	79	0.15	Tier 4

The dataset used in this research consists of 53,877 complaint reports categorized into 18 public service domains from LaporGub Central Java. Table II presents the numerical distribution of complaints across categories, while Figure 2 visualizes the same information to highlight the imbalance pattern among tiers. The data exhibit a pronounced skewness, with the top three categories Infrastruktur (17,935 samples, 33.3%), Sosial Masyarakat (9,074 samples, 16.8%), and Pendidikan (6,210 samples, 11.5%) collectively contributing over 60% of the total dataset. These classes, grouped within Tier 1 (>5,000 samples), dominate the dataset and represent public concerns that are infrastructural or community-related.

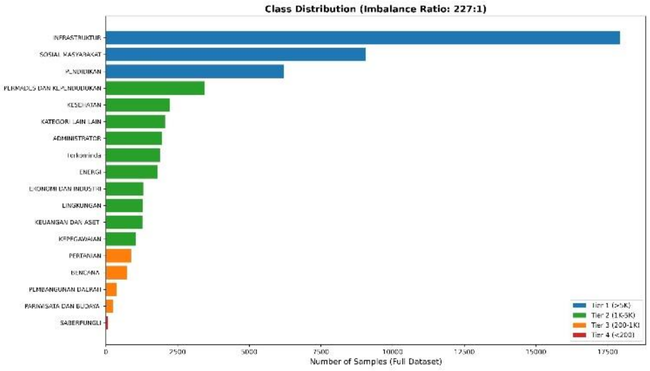

Figure 2. Class distribution visualization showing imbalance ratio (227:1) across tiers.

In contrast, lower-frequency categories such as Pertanian (878), Bencana (743), Pembangunan Daerah (380), And Pariwisata dan Budaya (248) fall into Tier 3 (200–1,000 samples), while Saberpungli constitutes an extreme minority (Tier 4) with only 79 samples (0.15%). This imbalance ratio of approximately 227:1 underscores the data's extreme disparity, which inherently biases model learning toward the dominant classes and constrains predictive reliability for underrepresented categories [15]. Such disproportionate representation also reflects real-world complaint behavior citizens tend to report physical and visible service issues (e.g., infrastructure, health, education) more frequently than administrative or governance concerns [1].

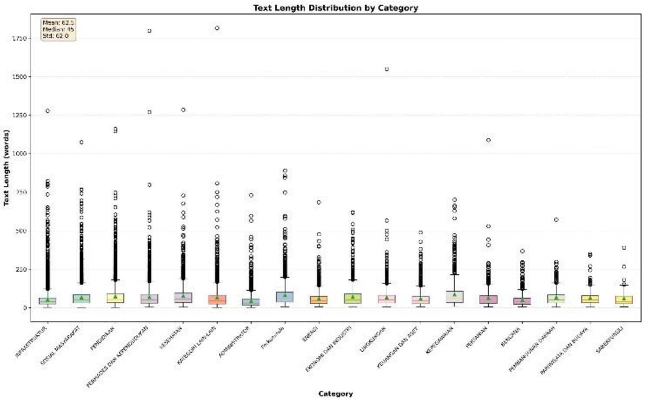

Figure 3. Text length distribution by category indicating short and variable complaint lengths.

Beyond label distribution, the dataset also varies linguistically. As shown in Figure 3, the average complaint length is approximately 62 words, with a median of 45 and standard deviation of 62, revealing substantial diversity in textual verbosity. Longer and more complex reports are common in categories such as Sosial Masyarakat, Pendidikan, and Kesehatan, while concise and formulaic submissions dominate categories like Pariwisata dan Budaya And Saberpungli. This diversity suggests that category complexity correlates with expression length and lexical richness, which can influence model performance by altering the amount of contextual information available for learning.

From a modeling perspective, these linguistic patterns justify the choice of 128 tokens as the maximum sequence length for transformer input long enough to capture the content of nearly all reports while minimizing excessive padding. Understanding these quantitative and textual imbalances provides the necessary context for interpreting model behavior and performance disparities in subsequent sections.

B. Overall Model Performance

After establishing the dataset characteristics, this subsection compares the overall performance of three classification models SVM + TF-IDF, XLM-RoBERTa (no class weighting), and XLM-RoBERTa with class weighting using macro-F1, weighted-F1, and accuracy as summarized in Table III.

TABLE III
OVERALL PERFORMANCE OF SVM AND XLM-R MODELS ACROSS 18 CATEGORIES

Model	Macro-F1	95% CI	Weighted-F1	Accuracy
SVM + TF-IDF	0.561	[0.544, 0.579]	0.737	0.734
XLM-R (no CW)	0.610	[0.595, 0.626]	0.780	0.792
XLM-R + CW	0.608	[0.594, 0.622]	0.767	0.760

The results reveal that XLM-RoBERTa without class weighting achieved the highest macro-F1 of 0.610, outperforming the traditional SVM + TF-IDF baseline (0.561). This improvement indicates that multilingual transformer representations are more capable of capturing contextual nuances in citizen complaints, particularly those involving complex expressions or mixed formal–informal phrasing [6]. The transformer's contextual embeddings contribute to more consistent class separation across categories, which explains the gain in macro-F1 despite similar overall accuracy levels.

Interestingly, applying class weighting to XLM-RoBERTa produced a slightly lower macro-F1 (0.608) than the unweighted variant but still within overlapping confidence intervals [0.594,0.622]. This suggests that class weighting neither improved nor significantly harmed overall performance. Instead, it likely shifted the model's focus from majority categories toward minority ones, a trade-off that increased sensitivity for rare classes while slightly reducing performance for frequent ones. Such marginal differences are consistent with findings in prior imbalance studies, where weighting stabilizes recall for minority categories but can reduce global precision [18].

The weighted-F1 and accuracy values further support this interpretation. While the SVM baseline achieved reasonable accuracy (0.734) and weighted-F1 (0.737), both XLM-RoBERTa variants scored higher accuracy 0.792 for the unweighted model and 0.760 for the class-weighted model demonstrating stronger overall generalization. The minor drop in accuracy after weighting reflects a deliberate redistribution of predictive attention rather than a true degradation of model quality. In practical terms, the weighted transformer model yields fairer class-level predictions, trading small reductions in overall accuracy for improved balance across categories.

Overall, these findings confirm that transformer-based architectures substantially outperform traditional vector-based models in handling semantically diverse and imbalanced complaint data. Moreover, the nuanced impact of class weighting underscores that while weighting may not always raise aggregate metrics, it remains valuable for promoting equity across underrepresented complaint categories a consideration explored further in the tier-based analysis that follows.

C. Tier-Based Analysis

To further investigate how class sample size influences model performance, we conduct a tier-based evaluation by aggregating per-class F1-scores within the four frequency tiers defined in Section II B (see Table IV and Figure 4). These tiers are not intended as universal rules; they summarise the specific long-tailed distribution of the LaporGub dataset, where three categories have more than 5,000 complaints, ten fall between 1,000 and 5,000, four lie between 200 and 1,000, and Saberpungli is the only category with fewer than 200 complaints. As a result, Tier 4 represents an extreme minority regime rather than a generic "small class" setting. The consistently near-zero F1-score for Saberpungli suggests that, in this dataset, classes with fewer than roughly 200 training examples are too rare for all three models to generalise reliably we therefore treat 200 as an empirical boundary between "rare but learnable" and "extremely rare" classes in the context of LaporGub.

TABLE IV
MEAN F1-SCORE PER CLASS SIZE TIER

Tier	n_classes	SVM	XLM-R	XLM-R + CW	Improvement
Tier 1	3	0.834	0.865	0.846	-2.2 %
Tier 2	10	0.564	0.633	0.632	-0.1 %
Tier 3	4	0.464	0.516	0.521	+0.9 %
Tier 4	1	0.100	0.000	0.000	-

Across all tiers, the results reveal a clear performance hierarchy in which XLM-RoBERTa-based models generally outperform the SVM baseline whenever there are at least a few hundred examples per tier, although the benefit of class weighting varies with class size. In Tier 1 (>5,000 samples), where abundant data provide sufficient representation, the SVM baseline already achieves strong performance (0.834 F1), but the unweighted XLM-RoBERTa model attains a higher mean F1 of 0.865. The class-weighted variant is slightly lower at 0.846 (Δ = -2.2 %), indicating that under well-represented conditions, class weighting is not strictly necessary and can marginally trade off performance on majority categories without delivering additional gains in aggregate tier-level F1.

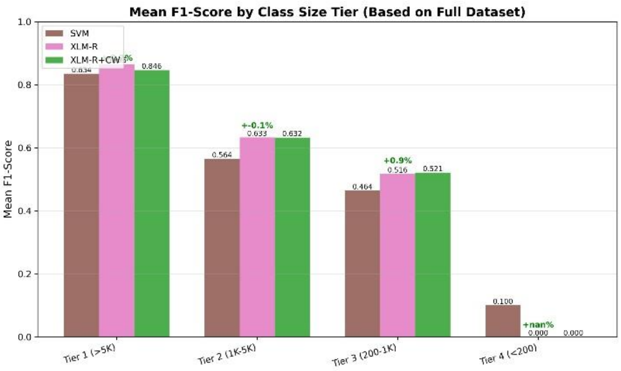

Figure 4. Mean F1-Score by Class Size Tier (Based on Full Dataset)

In Tier 2 (1,000-5,000 samples), the advantage of transformer-based models becomes more apparent. The mean F1 increases from 0.564 for SVM to 0.633 for the unweighted XLM-RoBERTa model, while the class-weighted configuration remains essentially identical at 0.632 (Δ ≈ –0.1 %). This pattern suggests that in medium-frequency regimes, the multilingual transformer already captures the dominant linguistic regularities, and class weighting primarily acts as a stabilising mechanism rather than a source of substantial additional improvement.

For Tier 3 (200-1,000 samples) the transition zone between minority and medium-frequency categories the benefit of class weighting is clearest. Mean F1 rises from 0.516 for the unweighted XLM-RoBERTa model to 0.521 for the class-weighted variant, corresponding to an absolute gain of +0.005 and a relative improvement of +0.9 %. Although numerically modest, this increase indicates that class weighting can partially compensate for data scarcity by placing additional emphasis on under-represented categories during optimisation. At the same time, the limited magnitude of the gain shows that bias is only partially mitigated and that lexical or contextual overlap between classes continues to constrain accuracy in this regime [15].

Finally, Tier 4 (<200 samples) represents the extreme minority regime, where all models struggle. The SVM baseline achieves a mean F1 of 0.100, whereas both transformer variants collapse to 0.000 F1, confirming that even with class weighting, the models fail to generalise when the number of training instances falls below roughly 200 examples the empirical threshold at which learning becomes unstable in this dataset [18]. This finding reinforces the earlier observation that algorithmic adjustments alone cannot substitute for adequate data representation.

Taken together, the tier-based results show that multilingual transformer models offer clear advantages over a traditional SVM baseline whenever each tier contains at least a few hundred samples (Tiers 1-3), while the additional contribution of class weighting is largely confined to the minority and transition regime in Tier 3, where it produces small but consistent gains and remains almost neutral in Tier 2 but slightly detrimental in Tiers 1 and 4. These patterns indicate that handling imbalance in large-scale complaint classification requires not only robust multilingual transformer architectures but also sufficient data density in rare categories, a theme that is further examined in the subsequent per-class analysis.

D. Per-Class Performance

To provide a finer-grained understanding of model behavior, Table V present the per-class F1-scores across all 18 complaint categories, arranged in descending order based on the proposed XLM-RoBERTa + Class Weighting (CW) model. This analysis examines both the best and worst-performing categories while interpreting how class size and linguistic complexity shape the observed outcomes.

TABLE V
F1-SCORE FOR ALL CATEGORIES

Category	**Test Samples**	**SVM**	**XLM-R**	**XLM-R + CW**	**Tier**
Infrastruktur	2691	0.869	0.898	0.877	Tier 1
Pendidikan	931	0.856	0.884	0.873	Tier 1
Kesehatan	337	0.754	0.786	0.808	Tier 2
Permades Dan Kependudukan	517	0.741	0.788	0.800	Tier 2
Energi	272	0.757	0.824	0.789	Tier 2
Sosial Masyarakat	1 361	0.778	0.813	0.788	Tier 1
Pertanian	132	0.767	0.818	0.784	Tier 3
Lingkungan	192	0.629	0.720	0.714	Tier 2
Keuangan Dan Aset	192	0.650	0.737	0.710	Tier 2
Ekonomi Dan Industri	195	0.606	0.737	0.688	Tier 2
Forkominda	283	0.547	0.652	0.613	Tier 2
Kepegawaian	157	0.544	0.555	0.610	Tier 2
Bencana	112	0.575	0.664	0.597	Tier 3
Pariwisata Dan Budaya	37	0.390	0.463	0.488	Tier 3
Administrator	294	0.220	0.317	0.341	Tier 2
Kategori Lain-Lain	310	0.194	0.212	0.249	Tier 2
Pembangunan Daerah	57	0.124	0.120	0.214	Tier 3
Saberpungli	12	0.100	0.000	0.000	Tier 4

Across categories, performance follows a clear correlation with class size. Major categories such as Infrastruktur (F1 = 0.877) and Pendidikan (F1 = 0.873) achieved the highest scores, reflecting the ample data available for these domains (Tier 1 > 5 K). The model's strong contextual representation allows it to capture diverse lexical patterns in these frequent complaint types. Similarly, Tier 2 categories with moderate sample sizes such as Kesehatan (F1 = 0.808) and Permades dan Kependudukan (F1 = 0.800) also show stable and high performance, indicating that around 1,000-5,000 samples provide a reliable foundation for effective generalization [18].

In contrast, categories belonging to the minority and extreme minority tiers show a steep decline in F1 performance. Classes like Administrator (F1 = 0.341), Pembangunan Daerah (F1 = 0.214), and Saberpungli (F1 = 0.000) represent substantial challenges for the model. Despite class weighting, Saberpungli remains unlearnable due to its extremely limited data (12 test samples) and high contextual overlap with administrative or anti-corruption reports. Similarly, the model confuses Pembangunan Daerah and Kategori lain-lain, which share similar lexical cues, resulting in low precision. These patterns reaffirm that while class weighting improves general fairness, it cannot fully overcome *data scarcity and semantic overlap* at the class level [15].

Interestingly, several mid-sized categories such as Pertanian (F1 = 0.784) and Lingkungan (F1 = 0.714) benefit from both balanced representation and distinct lexical boundaries, achieving consistent recognition. This indicates that category-specific clarity (semantic separability) contributes as much as sample size to classification reliability. Overall, the per-class trends show that linguistic ambiguity, rather than imbalance alone, plays a key role in misclassification for certain low-frequency domains.

These observations suggest that improving data coverage for ultra-minority classes and refining label definitions for semantically overlapping categories are necessary steps to enhance downstream model reliability an issue explored further in the following error analysis section.

E. Error Analysis

1) Confusion Patterns

To understand the model's weaknesses beyond aggregated scores, the most frequent misclassification patterns were examined using Table VI. The analysis reveals that errors predominantly occur among semantically related or operationally adjacent categories rather than being randomly distributed. A prominent example is complaints labelled Infrastruktur being misclassified as Bencana, where reports describing damaged roads, bridges, or public facilities are framed in the context of floods, landslides, or other disasters. In these cases, the model appears to prioritise disaster-related keywords over the underlying infrastructure maintenance aspect.

Another recurrent pattern involves Sosial Masyarakat being predicted as either Administrator or Kategori Lain-Lain. Complaints that mix community welfare issues with procedural or bureaucratic details invite ambiguity between social-policy and administrative handling, leading the classifier to favour more generic administrative labels when the social focus is not expressed explicitly. Similarly, Infrastruktur is frequently confused with Kategori Lain-Lain and, to a lesser extent, Sosial Masyarakat, particularly when the complaint text blends physical service issues with broader social impacts or vague descriptions of public facilities.

These patterns suggest that the classifier relies heavily on overlapping lexical fields such as references to public facilities, disasters, social assistance, and administrative processes whenever category boundaries are subtle. Quantitatively, the top-five confusion pairs in Table VI account for a substantial fraction of all misclassified instances, indicating that semantic proximity between categories is a stronger driver of error than class frequency alone [15].

TABLE VI
TOP-5 MOST CONFUSED CLASS PAIRS (XLM-R + CW)

True Label	Predicted Label	Count	True Class
Infrastruktur	Bencana	93	3.5 %
Sosial masyarakat	Administrator	80	5.9 %
Sosial masyarakat	Kategori lain-lain	74	5.4 %
Infrastruktur	Kategori lain-lain	69	2.6 %
Infrastruktur	Sosial masyarakat	55	2.0 %

2) Qualitative Error Examples

A qualitative inspection of misclassified samples, summarised in Table VII, further illustrates the underlying linguistic and contextual sources of error. In the broader error set, three recurring error types can be observed: semantic overlap, implicit language, and data scarcity. Semantic overlap arises when different categories share similar vocabulary or thematic content, making it difficult for the model to separate closely related domains. Implicit language occurs when the main topic of the complaint is not stated explicitly and must be inferred from indirect cues. Data scarcity, in contrast, reflects situations where very few training examples are available for a given category, limiting the model's ability to learn distinctive patterns.

Table VII primarily highlights cases of semantic overlap and data scarcity, which are the most prominent issues in the LaporGub dataset. One example of semantic overlap appears in a complaint about long-standing public dissatisfaction with the presence of a café in a residential area. The case is labelled as Pembangunan Daerah but predicted as Kategori Lain-Lain because the narrative mixes concerns about zoning, neighbourhood comfort, and local regulation, blurring the boundary between specific regional development policies and more generic administrative categories. A similar phenomenon occurs in disaster-related reports: a complaint describing ground cracks and landslide risk in a village is labelled as Bencana yet predicted as Infrastruktur, since the text emphasises damaged roads and physical facilities that closely resemble typical infrastructure-related complaints.

Data scarcity is clearly visible in minority categories such as Pembangunan Daerah and Pariwisata dan Budaya, as well as in several Bencana instances. For example, a report about high land and property tax (PBB) in a specific region is labelled as Pembangunan Daerah but predicted as Keuangan dan Aset, reflecting the model's tendency to gravitate towards more frequent, financially oriented categories when only a limited number of development-related training samples are available. Likewise, complaints involving tourism or cultural events are often mapped from Pariwisata dan Budaya to Kategori Lain-Lain when references to tourism or culture are brief and embedded within broader administrative concerns. Other disaster-related complaints, such as reports of toxic fumes from a fertiliser factory or nearly expired food aid for flood victims, are predicted as Lingkungan and Sosial Masyarakat respectively, again indicating that the model prefers more populated and semantically related labels under data-scarce conditions.

Beyond the specific examples listed in Table VII, additional misclassified cases (not shown for brevity) reveal instances of implicit language, where the true category is never mentioned explicitly but must be inferred from context. Together, these observations confirm that semantic ambiguity and limited data availability jointly drive many of the remaining errors, particularly for low-frequency and contextually entangled categories in the LaporGub complaint taxonomy.

TABLE VII
REPRESENTATIVE MISCLASSIFICATION EXAMPLES

Text	true label	predicted	Confidence	Error Type
“tetap tidak ada perubahaan warga sekitar sudah muak dengan keberadaan cafe xxx yang beralamat di...”	pembangunan daerah	kategori lain-lain	0.41	Semantic Overlap
“pak kenapa pajak pbb di xxx mahal daripd kabupaten lain?pdahal ka.....”	pembangunan daerah	keuangan dan aset	0.86	Data Scarcity
“jumat 11 november 2022, tepatnya di xxx,terjadi retakan tanah...”	bencana	infrastruktur	0.77	Semantic Overlap
“suatu kehormatan bagi pecinta merpati kolong, apa bila pak gubernur berke...”	pariwisata dan budaya	kategori lain-lain	0.57	Data Scarcity
“saay ingin melaporkan kabut berbau menyengat di daerah pucang gading yg di sebabkan oleh pembakaran pabrik pupuk saprotan kabutnya berbau menyengat da...”	bencana	lingkungan	0.97	Data Scarcity
“bantuan mie instan untuk korban rob pekalongan dari bpbd jawatengah hampir kadaluarsa 2 bulan lagi sangat mengecewakan...”	bencana	sosial masyarakat	0.88	Data Scarcity
“maaf,apakah bs mnt bantuan dr pak ganjar utk penyelesaian anggar...”	pariwisata dan budaya	kategori lain-lain	0.29	Data Scarcity
“saya melaporkan dengan no aduan lgmb45615100 blm di tanggapi sampai se...”	bencana	pembangunan daerah	0.21	Data Scarcity

3) Impact of Class Weighting and Interpretation

The comparison between XLM-RoBERTa models with and without class weighting shows a noticeable shift in the nature of errors. In the unweighted model, many minority-class complaints tend to be absorbed by a small set of dominant categories such as Infrastruktur or Sosial Masyarakat, indicating a strong bias in gradient updates under extreme imbalance. After applying class weighting, these errors increasingly manifest as semantic confusions between contextually related labels, as illustrated by the top confusion pairs in Table VI, rather than as systematic defaults to the most frequent classes. This pattern reflects progress toward more balanced learning across categories, even though the resulting discrimination is still imperfect.

However, extreme minority classes such as Saberpungli remain essentially unsolved, reinforcing the finding from the tier-based analysis that class weighting cannot compensate for data sparsity below roughly 200 training samples. In this regime, all evaluated models exhibit unstable F1-scores and near-zero recall, suggesting that algorithmic reweighting alone is insufficient and must be complemented by data-centric strategies such as targeted data augmentation, revised label definitions, or hierarchical label modelling [17].

Taken together, the error analysis indicates that while the class-weighted XLM-RoBERTa model improves fairness in the sense of distributing performance more evenly across low- and medium-frequency categories, residual misclassifications are still driven by semantic ambiguity and limited contextual diversity, especially in overlapping administrative and disaster-related domains. Addressing these issues through richer linguistic representations, additional training data for ultra-minority categories, and possibly task-specific label restructuring is crucial to achieving deployment-ready robustness in government complaint classification. From an operational perspective, reducing systematic misclassification of minority categories can help governments avoid overlooking sensitive cases and support more equitable service delivery. In a real deployment, more reliable routing for low- and medium-frequency complaint types would be expected to reduce manual triage workload and shorten response times for under-served cases, although quantifying such gains is beyond the scope of this research.

IV. CONCLUSION

This research provides empirical evidence that multilingual transformer models outperform traditional machine learning approaches in handling extreme class imbalance in Indonesian government complaint data. Among the evaluated models, the unweighted XLM-RoBERTa variant achieved the highest Macro-F1 of 0.610, compared to 0.561 for the SVM + TF-IDF baseline, while the class-weighted XLM-RoBERTa model attained a very similar

Macro-F1 of 0.608. These results show that contextual multilingual representations offer stronger overall generalisation than sparse vector baselines and that simple class weighting can redistribute performance towards minority categories without materially degrading aggregate Macro-F1, thereby supporting a fairer treatment of low- and medium-frequency complaint types in operational settings.

The tier-based and per-class analyses further revealed that class weighting is most beneficial for minority and medium-frequency categories, where it yields consistent but modest improvements in Macro-F1, while offering only limited gains for the extremely rare Saberpungli category with fewer than 200 complaints. In this regime, all models exhibit unstable F1-scores and near-zero recall, indicating that class weighting alone is insufficient when classes are extremely under-represented and that additional data-centric interventions are required. Error examination across all 18 categories shows that semantic overlap and implicit language dominate misclassifications, whereas data scarcity prevents the model from learning distinctive patterns for ultra-minority categories. Collectively, these findings highlight that label imbalance and linguistic complexity jointly constrain model performance in multilingual complaint classification, rather than data size alone, and they empirically extend prior work on fairness-aware class-imbalanced learning to an extreme 227:1 imbalance scenario in Indonesian e-government data.

Building on these insights, future research should pursue both data-centric and model-centric solutions, including targeted data augmentation for minority categories, focal loss or other cost-sensitive reweighting schemes for training stability, and hierarchical or multi-label classification strategies to reduce semantic overlap between related complaint types. The experimental framework and analyses presented here offer a reproducible reference point for subsequent studies in Indonesian NLP and e-governance analytics, and they position multilingual transformers with simple class weighting as a fair, scalable, and domain-adapted backbone for automated complaint routing in public service management systems.

ACKNOWLEDGMENT

The authors extend their profound appreciation to the Communication and Information Agency of Central Java Province (Dinas Komunikasi dan Informasi Provinsi Jawa Tengah) for granting access to the *Lapor Gub* complaint dataset, which served as a fundamental resource for this research. The authors also convey their gratitude to colleagues, mentors, and academic peers whose valuable suggestions and constructive feedback greatly contributed to refining this research. Furthermore, sincere thanks are directed to the supporting institutions for their continuous encouragement and provision of facilities that played an essential role in the successful completion of this research.

REFERENCES

[1] M. M. Priyadharshan, “A Digital Governance Framework for Intelligent Complaint Registration, Tracking, and Transparent Redressal,” vol. 8, no. 5, 2025.

[2] D. Hiremath, H. Patil, R. Patil, V. Hiremath, and C. R. Shivanagi, “Public Online Complaint Registration and Management System,” vol. 9, no. 3, 2024.

[3] F. Caldeira, L. Nunes, and R. Ribeiro, “Classification of Public Administration Complaints,” in *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*, J. Cordeiro, M. J. Pereira, N. F. Rodrigues, and S. Pais, Eds., in Open Access Series in Informatics (OASIcs), vol. 104. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022, p. 9:1-9:12. doi: 10.4230/OASIcs.SLATE.2022.9.

[4] T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” in *Machine Learning: ECML-98*, vol. 1398, C. Nédellec and C. Rouveirol, Eds., in Lecture Notes in Computer Science, vol. 1398. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142. doi: 10.1007/BFb0026683.

[5] Q. Li *et al.*, “A Survey on Text Classification: From Traditional to Deep Learning,” *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, pp. 1–41, Apr. 2022, doi: 10.1145/3495162.

[6] A. Conneau *et al.*, “Unsupervised Cross-lingual Representation Learning at Scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[7] M. I. Ragab, E. H. Mohamed, and W. Medhat, “Multilingual Propaganda Detection: Exploring Transformer-Based Models mBERT, XLM-RoBERTa, and mT5,” in *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, M. Jarrar, H. Habash, and M. El-Haj, Eds., Abu Dhabi: Association for Computational Linguistics, Jan. 2025, pp. 75–82. Accessed: Oct. 28, 2025. [Online]. Available: https://aclanthology.org/2025.nakbanlp-1.9/

[8] F. Ullah *et al.*, “Prompt-based fine-tuning with multilingual transformers for language-independent sentiment analysis,” *Sci Rep*, vol. 15, no. 1, p. 20834, Jul. 2025, doi: 10.1038/s41598-025-03559-7.

[9] C.-H. Lin and U. Nuha, “Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy,” *J Big Data*, vol. 10, no. 1, p. 88, May 2023, doi: 10.1186/s40537-023-00782-9.

[10] L. Afuan, N. Hidayat, H. Hamdani, H. Ismanto, B. C. Purnama, and D. I. Ramdhani, “Optimizing BERT Models with Fine-Tuning for Indonesian Twitter Sentiment Analysis,” *JOWUA*, vol. 16, no. 2, pp. 248–267, Jun. 2025, doi: 10.58346/JOWUA.2025.I2.016.

[11] W. Christian, D. Adamlu, A. Yu, and D. Suhartono, “Leveraging IndoBERT and DistilBERT for Indonesian Emotion Classification in E-Commerce Reviews,” Sep. 18, 2025, *arXiv*: arXiv:2509.14611. doi: 10.48550/arXiv.2509.14611.

[12] 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*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[13] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, K.-F. Wong, K. Knight, and H. Wu, Eds., Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[14] W. Wongso, D. S. Setiawan, S. Limcorn, and A. Joyoadikusumo, “NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural,” in *Proceedings of the Second Workshop in South East Asian Language Processing*, D. Wijaya, A. F. Aji, C. Vania, G. I.

Winata, and A. Purwarianti, Eds., Online: Association for Computational Linguistics, Jan. 2025, pp. 10–26. Accessed: Oct. 28, 2025. [Online]. Available: https://aclanthology.org/2025.sealp-1.2/

[15] S. Subramanian, A. Rahimi, T. Baldwin, T. Cohn, and L. Frermann, "Fairness-aware Class Imbalanced Learning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2045–2051. doi: 10.18653/v1/2021.emnlp-main.155.

[16] B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 9s, pp. 790–803, Jul. 2023.

[17] F. Boabang and S. A. Gyamerah, "An Enhanced Focal Loss Function to Mitigate Class Imbalance in Auto Insurance Fraud Detection with Explainable AI," Aug. 04, 2025, *arXiv*: arXiv:2508.02283. doi: 10.48550/arXiv.2508.02283.

[18] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples".

[19] S. m T. I. Tonmoy, "Embeddings at BLP-2023 Task 2: Optimizing Fine-Tuned Transformers with Cost-Sensitive Learning for Multiclass Sentiment Analysis," in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, F. Alam, S. Kar, S. A. Chowdhury, F. Sadeque, and R. Amin, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 340–346. doi: 10.18653/v1/2023.banglalp-1.46.

[20] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[21] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci Rep*, vol. 15, no. 1, p. 21631, Jul. 2025, doi: 10.1038/s41598-025-05791-7.

[22] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization.," 2019.