STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–28.
Published online in International Academic Press (www.IAPress.org)

A Systematic Comparison of Cost-Sensitive Methods for Long-Tail Intent Classification in Indonesian Government Complaint Routing

Dewi Agustini Santoso[1*], Heni Indrayani[2], Karis Widyatmoko[1], Farrikh Alzami[3], Iswahyudi[4]

[1]*Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
[2]*Department of Communication Science, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
[3]*Department of Information Systems, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
[4]*Communication and Information Agency of Central Java Province, Semarang, Indonesia*

Abstract E-government complaint routing systems face extreme class imbalance under which standard transformer fine-tuning leaves most minority agencies without correct routings. This study compares two cost-sensitive loss modification methods, inverse-frequency class weighting and focal loss, for long-tail intent classification in Indonesian government complaint routing. The dataset comprises 56,068 Indonesian-Javanese-English code-mixed complaints across 86 government agencies with an imbalance ratio of 332.86:1, among the most extreme long-tail distributions documented in e-government NLP research. XLM-RoBERTa-base was fine-tuned under three configurations: a CrossEntropyLoss baseline, inverse-frequency class weighting, and focal loss with a focusing parameter of 2.0. Evaluation employed Pareto-based head/tail segmentation, where 33 head classes account for 80% of training data and 53 tail classes cover the remaining 20%. Experiments were repeated across three independent random seeds to assess stability. Class weighting improved tail macro F1 from 0.2048 to 0.3426, a 67.3% gain, with a corresponding 9.4% decline in head macro F1. Focal loss yielded a 63.7% tail gain but incurred a 15.4% head performance reduction. A Friedman test comparing macro F1 across three methods and three seeds yields a test statistic of 6.00 with two degrees of freedom and a p-value of 0.050, a boundary result that reflects perfectly consistent method ranking across seeds and is interpreted as directional evidence rather than conclusive proof. Segment-level Wilcoxon tests on per-class F1 are individually significant in every seed, for both the tail improvement and the head decline, with p-values below 0.001 in all cases, indicating that the non-significant pooled test conceals two opposing within-segment effects. Class weighting demonstrated superior balance, achieving a higher overall macro F1 (0.4441) and higher tail F1 than focal loss (0.4241) under identical training conditions. Spearman rank correlation between class training frequency and F1 improvement is consistently negative across all seeds (mean correlation of -0.585, p-value below 0.001), confirming that lower-frequency agencies benefit systematically from cost-sensitive training. The Pareto-based evaluation convention adopted here provides a reproducible, threshold-free decomposition of imbalanced classifier performance that transfers to other domains with majority/minority class structure.

Keywords Cost-sensitive learning, Long-tail classification, Indonesian NLP, Government complaint routing, XLM-RoBERTa, Focal loss, Class imbalance

AMS 2010 subject classifications 68T50, 62H30

DOI: 10.19139/soic-2310-5070-4091

1. Introduction

Indonesian provincial governments receive over 1,000 citizen complaints daily through online platforms. The Jateng Ngopeni Nglakoni system serves Central Java Province by routing these complaints to government agencies

*Correspondence to: Dewi Agustini Santoso (Email: dewi@dsn.dinus.ac.id). Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Jl. Imam Bonjol No. 207, Semarang 50131, Indonesia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

automatically. Manual routing is infeasible at this operational scale, and the accuracy of automated routing determines whether citizens receive timely responses from the agencies responsible for their concerns.

Complaint distribution follows a power law. Approximately 80% of submitted complaints are directed toward fewer than 40% of registered agencies, primarily infrastructure and public works departments. Tail classes serve specialised functions, including anti-corruption bureaus, regional hospitals, and specific local programmes, and receive fewer than 10 training samples in many cases. Standard cross-entropy fine-tuning of transformer models achieves high aggregate accuracy by concentrating predictive capacity on majority agencies while effectively ignoring minority ones.

The practical consequences of this failure mode are severe. Citizens with legitimate complaints directed at minority agencies experience systematic misrouting, producing delays and unresolved service requests. This inequity violates the e-government principle of equal access to public services [1]. The problem is compounded by the linguistic composition of complaints: Indonesian citizens routinely employ code-mixed text combining formal Indonesian, colloquial Indonesian, Javanese, and English, a setting in which monolingual models struggle [2].

Cost-sensitive learning addresses class imbalance at the loss function level without modifying training data. Two principal variants exist. Inverse-frequency class weighting assigns higher loss penalties to minority classes, making gradient updates proportional to class rarity. Focal loss [3] dynamically reduces the contribution of easy, well-classified examples while concentrating training on hard instances. Both approaches have been applied to vision tasks and English NLP, and focal loss has recently been applied to Indonesian government complaints for topic classification [4]. What remains missing is a systematic, statistically validated comparison of the two cost-sensitive variants under identical multilingual transformer configurations at the imbalance severity characteristic of operational agency routing.

1.1. Research Gap

Three gaps in existing literature motivate this study. Henning et al. [5] identified that no standardised evaluation protocol exists for imbalanced NLP and that multilingual settings remain almost entirely unexplored. Cost-sensitive training for transformer-based text classification has been demonstrated for English [6] but not for low-resource multilingual settings with code-mixing. No direct comparison of class weighting and focal loss under identical conditions exists in the e-government NLP domain; the closest prior work applies a single cost-sensitive method without a competing variant [4]. E-government text classification research typically addresses around 20 or fewer categories, with imbalance ratios that are either unreported or well below the 300:1 regime, leaving the behaviour of cost-sensitive methods under such conditions uncharacterised.

This study addresses three research questions. RQ1: Can cost-sensitive loss functions improve tail class performance without degrading head class performance below operational thresholds? RQ2: Does class weighting or focal loss provide superior head/tail balance under extreme long-tail conditions? RQ3: Is the performance improvement of cost-sensitive methods stable across multiple random seeds and of practical operational magnitude?

1.2. Contributions

The central motivation of this work is operational service equity: every misrouted complaint delays a citizen's access to the agency responsible for their concern, and misrouting concentrates precisely on the minority agencies that serve specialised needs. The study is therefore positioned as an empirical comparison that provides deployment-relevant evidence for a specific, severe, and previously uncharacterised operating regime, rather than as a new learning algorithm. This study makes four explicit contributions.

1. A corpus of 56,068 Indonesian-Javanese-English code-mixed government complaints across 86 agencies is characterised, the largest documented Indonesian complaint corpus with an imbalance ratio of 332.86:1, establishing a reference point for extreme long-tail e-government text classification.
2. A Pareto-based head/tail segmentation is adopted as an evaluation convention for threshold-free, data-driven decomposition of imbalanced classifier performance into majority and minority class contributions, enabling reproducible comparison across datasets and methods.

3. A systematic comparison of class weighting and focal loss under identical training conditions, evaluated across three independent random seeds, is conducted for multilingual transformer fine-tuning, with statistical validation via the Friedman test, pooled and segment-level Wilcoxon tests, and bootstrap confidence intervals over classes.
4. Quantified evidence is provided that class weighting attains higher overall macro F1 and tail F1 than focal loss under the 332.86:1 imbalance regime, together with an operational estimate of the routing improvement and its assumptions for the deployed e-government system.

2. Related Work

2.1. Class Imbalance in NLP

Approaches to class imbalance fall into three levels. Data-level techniques modify the training distribution through re-sampling or augmentation. Loss-level techniques modify the objective function while leaving the data unchanged, as in class weighting and focal loss. Representation-level techniques reshape the embedding space so that minority classes become separable. The present study operates at the loss level; this subsection reviews all three levels to position that choice. Addressing class imbalance in NLP has attracted growing attention as transformer models are deployed in production classification systems. Buda et al. [7] conducted a systematic study of class imbalance in convolutional networks, establishing that re-weighting and re-sampling strategies produce complementary benefits across a range of imbalance ratios. Huang et al. [8] extended this analysis to multi-label text classification with long-tailed distributions, demonstrating that label co-occurrence patterns complicate standard re-balancing and that no single technique dominates across all data conditions. Henning et al. [5] surveyed 14 methods across six datasets, identifying the absence of standardised evaluation protocols for imbalanced NLP as a critical gap and noting that multilingual settings remain nearly unexplored. Taskiran et al. [9] benchmarked 30 SMOTE variants on transformer-embedded text, confirming through Friedman testing that oversampling benefits are dataset- and classifier-dependent, with no single oversampling technique consistently superior.

Loss modification represents a complementary approach to data-level techniques. Lin et al. [3] introduced focal loss for dense object detection. The formulation $FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$ dynamically down-weights well-classified examples, concentrating training on hard instances. Li et al. [10] proposed Dice loss for imbalanced NLP tasks, demonstrating gains on named entity recognition and machine reading comprehension. Fernando and Tsokos [11] developed dynamically weighted balanced loss, combining confidence calibration with class-specific penalties. Cao et al. [12] proposed label-distribution-aware margin (LDAM) loss motivated by minimising a generalisation bound. Ridnik et al. [13] introduced asymmetric loss for multi-label classification, decoupling the treatment of positive and negative samples to address gradient imbalance in extreme many-label settings. Zhang et al. [14] surveyed deep long-tailed learning, noting that NLP applications lag substantially behind computer vision in systematic evaluation of these methods. Hu et al. [15] proposed weighted difference loss for multi-label classification, combining inter-class margin penalties with label-frequency adjustments and showing gains over focal variants on sparse-label benchmarks. Data-level and loss-level treatments are also combined in practice: Abdulkareem and Abdulazeez [16] paired generative oversampling with focal loss for imbalanced multi-class medical classification, reporting macro F1 gains on minority classes, and Winarno et al. [17] proposed selecting the loss function adaptively from the observed imbalance ratio, applying focal loss above a severity threshold and weighted cross-entropy below it. A caveat applies when transferring these findings from computer vision to NLP: vision benchmarks typically involve continuous input spaces and hundreds of examples per tail class, whereas long-tail text classification confronts discrete token distributions, semantic overlap between class vocabularies, and tail classes with single-digit sample counts, so relative method rankings established on vision benchmarks require re-validation on text.

Beyond loss modification, representation-level approaches have been explored. Xu et al. [18] proposed label-specific feature augmentation for long-tailed multi-label text, transferring intra-class semantic variation from head to tail labels in latent space without external data. Khalid et al. [19] applied label-supervised contrastive

learning in Euclidean and hyperbolic embedding spaces for imbalanced text classification, improving minority-class separability over standard cross-entropy baselines. Khvatskii et al. [20] combined class-aware contrastive optimisation with denoising autoencoding in the CAROL framework, achieving minority-class gains across diverse text benchmarks. Agbesi et al. [21] proposed dual-head alternating attention with adaptive convolution for low-resource and imbalanced classification, exploiting class-conditional attention to redistribute gradient signal toward rare categories. Mahmoudi and Salem [22] developed BalBERT, which integrates dataset-balancing procedures directly within the BERT fine-tuning loop, showing improvements over decoupled re-sampling on standard benchmarks. None of these methods has been evaluated under the extreme long-tail conditions characteristic of operational government complaint routing, nor compared systematically against one another under identical multilingual transformer configurations.

2.2. Indonesian and Multilingual NLP

Indonesian NLP infrastructure has expanded considerably. Koto et al. [23] released IndoLEM alongside IndoBERT, a pre-trained BERT model achieving competitive performance on seven downstream tasks. Wilie et al. [24] introduced IndoNLU covering 12 tasks spanning sentiment analysis, named entity recognition, and natural language inference. Conneau et al. [25] developed XLM-RoBERTa for 100 languages, achieving a 14.6% improvement over mBERT on cross-lingual benchmarks. The RoBERTa family has proven effective for specialised text classification beyond general-purpose benchmarks; Almansour et al. [26] reported that a fine-tuned RoBERTa encoder captures contextual payload patterns in cross-site scripting detection without manual feature engineering, illustrating the transfer of contextual representations to domain-specific classification tasks. Winata et al. [27] released NusaX, a parallel sentiment dataset covering 10 Indonesian local languages, demonstrating that regional language variation is a persistent challenge for multilingual models fine-tuned on standard Indonesian corpora.

Code-mixing poses additional challenges. Hidayatullah et al. [2] created the first labelled corpus for Indonesian-Javanese-English code-mixed text at the word level, providing the linguistic foundation for the present dataset. Hidayatullah et al. [28] developed a pre-trained language model for this trilingual code-mixed setting, confirming that domain-matched pre-training reduces token-level misidentification in code-mixed sequences. Hidayatullah et al. [29] further showed that transformer models substantially outperform recurrent baselines on code-mixed sentiment tasks in this trilingual setting. Adilazuarda et al. [30] evaluated robustness of Indonesian transformers against diverse code-mixed local languages, finding that performance degrades substantially under unseen code-mixing patterns, underscoring the importance of domain-specific evaluation. Wiciaputra et al. [31] demonstrated that XLM-RoBERTa transfers effectively to Indonesian without task-specific pre-training.

Beyond code-mixing, class imbalance in Indonesian NLP has received targeted attention. Cahya et al. [32] applied SMOTE and augmentation with IndoBERT to imbalanced multi-label Indonesian classification, achieving improvements on minority labels but limited to binary label settings with modest imbalance ratios. Nabiilah and Suhartono [33] combined three Indonesian pre-trained transformer variants through majority-vote ensemble with oversampling for personality classification on a heavily imbalanced Twitter dataset, achieving a macro F1 gain of 4.5 points over the best single model. Hutama and Suhartono [34] demonstrated that combining XLM-RoBERTa with BERTopic topic modelling improves classification accuracy on Indonesian hoax news, illustrating the utility of multilingual transformers for Indonesian information processing beyond sentiment analysis. Findawati et al. [35] addressed multi-label imbalanced dangerous speech classification on Indonesian government-related Twitter content, proposing a keyword-driven ensemble that handles label overlap and data scarcity. Riyadi et al. [36] introduced a domain-specific model for Indonesian government documents, confirming the advantage of targeted pre-training for formal administrative text.

Despite these advances, the intersection of code-mixed Indonesian, extreme long-tail imbalance, and operational agency routing has not been studied. Existing work either addresses Indonesian text without imbalance or addresses imbalance without the severe distributional conditions of government routing systems.

2.3. E-Government Complaint Classification

Table 1 positions this study within the e-government complaint classification literature; entries marked n.r. correspond to studies that do not report a numeric imbalance ratio. Ngai et al. [46] conducted a comprehensive

Table 1. Summary of related work on government complaint classification

Study	Language	Classes	Ratio	Method
Kim & Hong [37]	English	21→14	n.r.	Re-sampling
Tang et al. [38]	Chinese	16	≈130:1	Augmentation
Hashemi [39]	English	14	Moderate	Deep learning
Intani et al. [40]	Indonesian	20	Moderate	Deep learning
Madyatmadja et al. [41]	Indonesian	–	–	Contextual NLP
Wang et al. [42]	Chinese	–	Moderate	Joint attention
Xiong et al. [43]	Chinese	22	Moderate	BERT-BiLSTM-CNN
Zahro et al. [4]	Indonesian	18	n.r.	Focal loss
Esperanca et al. [44]	Portuguese	4	Moderate	Clustering
Chen et al. [45]	Chinese	–	–	Label correction
This study	Indonesian	86	332.86:1	Loss modification

literature review of NLP in government applications, finding that most deployments address sentiment analysis, topic modelling, and policy document processing, while operational routing to organisational units under severe class imbalance remains absent from the literature. Kim and Hong [37] classified Boston 311 transportation requests using a CNN with SMOTE-ENN oversampling on 10,000 training and 6,000 test samples; the original 21 categories were merged into 14 through peer evaluation, and no numeric imbalance ratio is reported. Tang et al. [38] handled 3,438 Chinese electricity complaints across 16 secondary categories through BERT with back-translation and synonym-substitution augmentation; the published category counts range from 1,434 to 11 samples, an imbalance ratio of approximately 130:1. Neither study examines agency routing under severe long-tail conditions.

More recent work has extended the domain. Hashemi [39] applied deep learning to 311 service request classification across 14 categories, confirming that text-based features outperform structured metadata alone but without addressing long-tail imbalance. Intani et al. [40] automated public complaint classification through the JakLapor system in Jakarta, Indonesia, across 20 categories using transformer-based models, but the study reports only aggregate accuracy and does not address minority class failure modes. Madyatmadja et al. [41] developed a contextual text analytics framework for citizen report classification in Indonesian, demonstrating improvements over keyword-based baselines without examining cross-agency routing fairness. Wang et al. [42] proposed a joint attention enhancement network incorporating external keyword information for citizen complaint classification in Chinese, reducing noise sensitivity in long texts but without addressing class imbalance. Xiong et al. [43] applied BERT-BiLSTM-CNN feature fusion to public service request classification across 22 Chinese categories, achieving strong aggregate performance but again without minority-class decomposition.

Zahro et al. [4] constitute the most direct prior work. They fine-tuned XLM-RoBERTa with focal loss on 53,774 Indonesian complaints from the LaporGub channel of the same provincial ecosystem, classifying complaints into 18 topic categories and raising macro F1 from 0.606 to 0.625 relative to a cross-entropy baseline. The present study differs in three respects. First, the target is agency routing across 86 organisational units rather than topic classification across 18 categories; routing determines which office receives the complaint and exhibits a more severe long-tail because administrative concentration places most citizen interactions within a small number of offices. Second, this study compares two cost-sensitive variants under identical conditions with statistical testing, whereas Zahro et al. evaluate focal loss alone. Third, this study decomposes performance into head and tail segments, a breakdown absent from that work, which reports only aggregate metrics. Esperanca et al. [44] developed semantic clustering for Portuguese complaints, which does not address production routing to labelled organisational units. Chen et al. [45] focused on complaint resolution time prediction rather than routing. Jiang et al. [1] conducted a systematic review finding that most government NLP datasets remain proprietary, limiting reproducibility.

Across the studies in Table 1, category counts remain at or below 22, imbalance ratios are either unreported or no higher than approximately 130:1, and data manipulation dominates over loss modification. The present study addresses 86 agencies with a 332.86:1 ratio, the most extreme documented among these comparison studies, and

employs loss-function approaches that leave the training data unchanged. None of the compared studies reports a head/tail decomposition of performance, which prevents direct quantitative comparison of minority-segment gains and motivates the evaluation convention adopted here.

3. Methodology

Figure 1 presents the overall research framework, progressing through five sequential stages: raw data collection, preprocessing and quality filtering, Pareto-based class segmentation, cost-sensitive fine-tuning of XLM-RoBERTa under three experimental configurations, and evaluation using head/tail-stratified metrics with statistical significance testing.

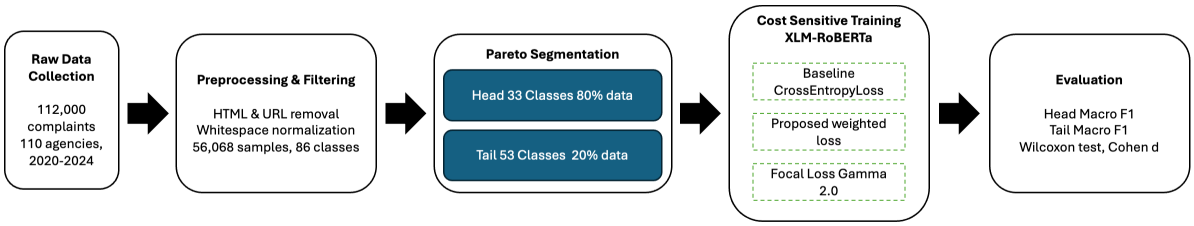


Figure 1. Overview of the research framework spanning data collection through evaluation.

The framework in Figure 1 isolates the loss function as the sole experimental variable. All other components, including data preprocessing, tokenisation, model architecture, training hyperparameters, and evaluation protocol, are held constant across the three experimental conditions.

3.1. Dataset

The Jateng Ngopeni Nglakoni platform (lakon.jatengprov.go.id) serves as the primary citizen complaint system for Central Java Province. Data collection spanned 2020 through 2024, yielding four years of operational records from 110 registered government agencies. The original corpus contained over 112,000 complaints.

Preprocessing applied three sequential operations. HTML artifacts, including tags such as `<br>`, ` `, and `<p>`, were removed. URLs containing `http` or `https` were eliminated. Multiple whitespace characters were collapsed to a single space. Emojis and punctuation were deliberately preserved, as both carry sentiment and contextual signals relevant to agency identification.

Quality filters removed three categories of problematic entries: texts reduced to empty strings after preprocessing, records with missing ground-truth agency assignments, and agencies with fewer than 10 training samples. The minimum-samples threshold ensures that stratified splitting across train, validation, and test partitions produces at least one sample per agency in each split. After filtering, 24 agencies (21.8% of the original 110) were excluded, resulting in a loss of approximately 2,156 complaints (1.9% of the pre-filtered corpus). The evaluated model consequently covers 86 of the 110 registered agencies; the excluded ultra-rare agencies remain outside the scope of the reported results and are revisited in the limitations. Final statistics are presented in Table 2.

Table 2 reflects the extreme frequency skew inherent in operational government complaint data. The most frequent agency accumulates 2,330 training samples while the rarest retains only 7. Code-mixed text appears throughout, with formal Indonesian dominating alongside Javanese honorifics and occasional English technical terminology. A representative example reads: “Dalan rusak parah di gang belakang rumahku, mbok tolong diperbaiki,” translating as “Road severely damaged behind my house, please fix.” The complete list of 86 agencies with their Pareto segment assignments and test set sample counts is provided in Appendix A.

The corpus derives from operational citizen complaint records held by the Communication and Information Agency of Central Java Province (Diskominfo Jawa Tengah). Because individual complaints contain personally identifiable citizen information, applicable data protection regulations prohibit public release of the records in any form, and the dataset therefore cannot be deposited in an open repository. Qualified researchers may direct

Table 2. Dataset statistics summary

Attribute	Value
Total samples	56,068
Training samples (70%)	39,246
Validation samples (15%)	8,411
Test samples (15%)	8,411
Number of agencies (classes)	86
Head classes (Pareto 80%)	33
Tail classes (Pareto 20%)	53
Minimum training frequency	7
Maximum training frequency	2,330
Imbalance ratio	332.86:1
Mean training frequency	456.35
Median training frequency	326.50

access enquiries to Diskominfo Jawa Tengah, which evaluates requests under its data governance procedures. To support verification despite this constraint, the complete experimental pipeline, including preprocessing, training, evaluation, and statistical testing code, is available at https://github.com/dsnalzami/jnn2026, together with the aggregate per-class metric tables from which every reported figure and statistic can be reproduced.

3.2. Pareto-Based Head/Tail Segmentation

Prior studies employ arbitrary thresholds to define majority and minority classes, such as "tail equals classes with fewer than 100 samples," without principled justification. Threshold choice substantially affects evaluation conclusions. The Pareto Principle provides an objective, data-driven alternative [14]. Classes are sorted by descending training frequency. A cumulative coverage percentage is computed. Head classes are defined as the minimal set covering exactly 80% of training samples; the remaining classes constitute the tail, accounting for 20% of samples. Figure 2 illustrates this segmentation on the present dataset.

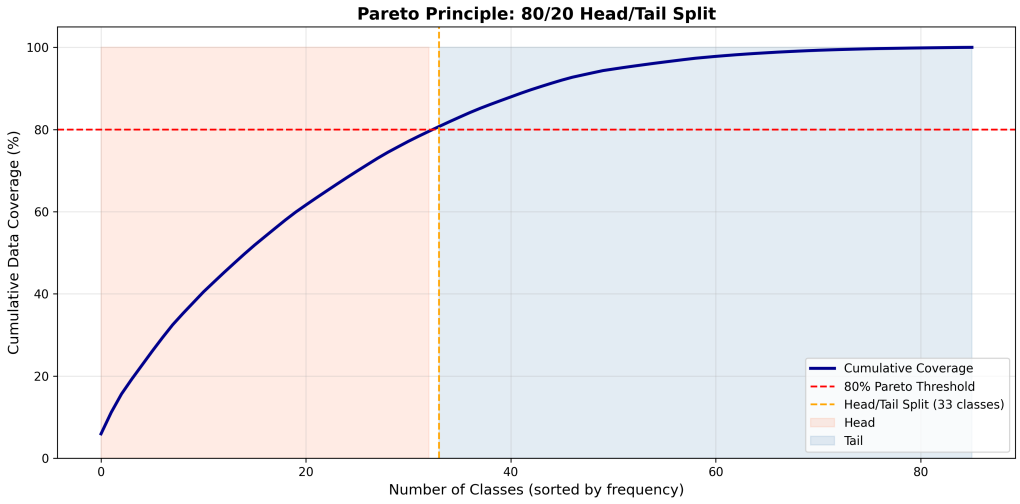


Figure 2. Cumulative training data coverage by class rank. The vertical line marks the 80% threshold at 33 head classes.

Figure 2 confirms the long-tail structure: 33 agencies (38.4% of classes) account for 80% of training data, while 53 agencies (61.6% of classes) contribute only 20%. This segmentation is computed exclusively on training frequencies to prevent leakage of test distribution information. Evaluation then computes macro F1 separately for each segment, making head/tail trade-offs explicit and reproducible.

Input: Training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, class counts $\{n_k\}_{k=1}^{K}$, pre-trained model f_θ, method $m \in \{\text{baseline, weight, focal}\}$, focal parameter $\gamma = 2.0$, training epochs $E = 4$
Output: Fine-tuned model f_θ^*, test predictions $\hat{y}$

Compute class weights $w_k \leftarrow N/(K \times n_k)$ for $k = 1, \ldots, K$;
Sort classes by descending frequency; compute cumulative coverage c_k;
$\mathcal{H} \leftarrow \{k : c_k \leq 0.80\}, \quad \mathcal{T} \leftarrow \{k : c_k > 0.80\}$;

for *epoch* $e = 1$ **to** E **do**
  for *each mini-batch* $(\mathbf{x}, \mathbf{y})$ **do**
    $\mathbf{l} \leftarrow f_\theta(\mathbf{x})$; `// forward pass`
    $p_t \leftarrow \text{softmax}(\mathbf{l})_y$; `// true-class probability`
    if $m =$ *baseline* **then**
      $\mathcal{L} \leftarrow -\log(p_t)$;
    else if $m =$ *weight* **then**
      $\mathcal{L} \leftarrow -w_y \cdot \log(p_t)$;
    else if $m =$ *focal* **then**
      $\mathcal{L} \leftarrow -w_y \cdot (1 - p_t)^\gamma \cdot \log(p_t)$;
    end
    Update θ via AdamW with $\nabla_\theta \mathcal{L}$;
  end
  Evaluate validation macro F1; save checkpoint if improved;
end

Load best checkpoint f_θ^*;
$\hat{y} \leftarrow \arg\max f_\theta^*(\mathbf{x}_{\text{test}})$;
return $f_\theta^*, \hat{y}$;

Algorithm 1: Cost-Sensitive Training for Long-Tail Complaint Classification

3.3. Model Architecture

XLM-RoBERTa-base provides the pre-trained foundation [25]. The model was pre-trained on CommonCrawl data from 100 languages, including Indonesian. It handles code-mixed Indonesian-Javanese-English text without requiring separate language identification at inference time. The architecture comprises 12 transformer encoder layers with 768-dimensional hidden representations and 12 attention heads, totalling approximately 270 million parameters. Tokenisation employs the SentencePiece tokeniser with a 250,000-entry vocabulary. Maximum sequence length was set to 128 tokens; analysis of the complaint corpus confirmed that 95% of texts fall within this limit after preprocessing. The classification head appends a linear projection from 768 dimensions to 86 output logits, with dropout at 0.1 following standard fine-tuning practice.

3.4. Cost-Sensitive Training Methods

Three experimental configurations were evaluated under identical conditions to isolate the effect of the loss function. Algorithm 1 presents the training pseudocode.

The baseline configuration applies standard CrossEntropyLoss with no modification. All classes receive equal gradient weight, expressed as $\mathcal{L} = -\log(p_t)$, where p_t is the probability assigned to the true class. This condition represents the current operational default for transformer fine-tuning and serves as the control against which both cost-sensitive methods are evaluated.

The class weighting method assigns per-class penalties through balanced weighting. Weights are computed using the scikit-learn balanced formula: $w_k = N/(K \times n_k)$, where N denotes total training samples (39,246), K the number of classes (86), and n_k the sample count for class k. The resulting weight range spans from 0.196 to 65.388, a 334-fold difference that mirrors the data imbalance. The modified loss becomes $\mathcal{L} = -w_y \cdot \log(p_t)$,

where w_y is the weight for true class y. A custom Trainer subclass overrides the `compute_loss` method with weights transferred to GPU at runtime. Model architecture remains unchanged.

Focal loss introduces dynamic difficulty-weighting on top of class weighting [3]. The formulation reads: $FL(p_t) = -w_y \cdot (1 - p_t)^\gamma \cdot \log(p_t)$. The focusing parameter $\gamma = 2.0$ follows the original paper recommendation. The alpha term reuses the inverse-frequency class weights from the class weighting method, creating a controlled comparison where the additional variable is exclusively the $(1 - p_t)^\gamma$ modulation. Easy examples with high p_t contribute less to the loss; hard examples with low p_t receive amplified gradients. A separate custom Trainer implements focal loss using log-softmax for numerical stability. This design entails a deliberate trade-off: because the focal condition inherits the alpha weights, it measures the incremental effect of difficulty modulation on top of frequency weighting rather than the effect of pure focal loss without alpha weighting. An unweighted focal condition would isolate the modulation term in absolute terms and is identified as an ablation for future work.

3.5. Training Configuration and Evaluation

Training configuration was identical across all three experimental conditions. Table 3 lists the full settings.

Table 3. Training hyperparameters, fixed across all experiments and seeds

Parameter	**Value**
Base model	xlm-roberta-base
Max sequence length	128 tokens
Batch size (per device)	8
Gradient accumulation	4 steps
Effective batch size	32
Learning rate	2×10^{-5}
Optimizer	AdamW
Weight decay	0.01
Warmup ratio	0.1
Epochs	4
Mixed precision	FP16
Focal γ	2.0
Random seeds	42, 123, 2024

Table 3 confirms that all experimental differences are attributable solely to the loss function. Linear warmup covered 10% of total training steps. Gradient accumulation over four steps yielded an effective batch size of 32. Early stopping monitored validation macro F1; the checkpoint achieving the highest validation score was loaded for final test set evaluation. Mixed precision (FP16) and gradient checkpointing permitted batch size 8 on a 16 GB GPU.

The primary evaluation metric is macro F1, computed as Macro F1 $= (1/K) \sum_{i=1}^{K} F1_i$, which assigns equal weight to all 86 agencies regardless of frequency. This metric is appropriate for extreme imbalance because it prevents high-frequency classes from dominating aggregate scores. Secondary metrics include weighted F1, overall accuracy, macro precision, and macro recall. Head macro F1 and tail macro F1 quantify the head/tail trade-off directly.

Statistical testing employed the Wilcoxon signed-rank test to compare per-class F1 distributions between baseline and class weighting. Normality of the paired differences was assessed via the Shapiro-Wilk test as a preliminary step for test selection between a parametric and non-parametric comparison, with the outcome of this diagnostic reported together with the significance results in Section 5. The one-tailed alternative (H_1: baseline F1 $<$ class-weighted F1) reflects the directional hypothesis motivated by the inverse-frequency weighting mechanism. Effect size is quantified by the standard paired Cohen's d, defined as mean difference divided by standard deviation of differences (ddof=1). To assess stability across data partitions and initialisation conditions, the full experimental pipeline was repeated under three independent random seeds (42, 123, 2024). Each seed controlled data splitting and model initialisation, producing a genuinely distinct experimental condition. Cross-seed comparison used the Friedman test on macro F1 values across three methods and three seeds. Because the

head and tail segments are hypothesised to move in opposite directions, two-sided Wilcoxon signed-rank tests were additionally applied to the per-class F1 differences within each segment separately, per seed. Uncertainty in the segment-level macro F1 differences was quantified with nonparametric bootstrap confidence intervals, resampling classes with replacement (10,000 resamples, percentile method), and seed-level 95% confidence intervals for overall macro F1 were computed with the t distribution on two degrees of freedom.

4. Experimental Setup

Experiments ran on Kaggle notebooks using dual NVIDIA T4 GPUs with 16 GB memory each. Training time approximated 4.4 hours per complete run (three experiments across four epochs). The software stack comprised Python 3.10, PyTorch 2.0, HuggingFace Transformers 4.30, and scikit-learn 1.3. Three independent runs were executed across separate Kaggle accounts to allow full GPU quota utilisation for each seed. Account separation affects only quota allocation; each run executed the identical notebook with the same pinned software versions, the same GPU model, and deterministic seeding of data splitting and model initialisation, so the execution environment was uniform across runs up to hardware scheduling that is outside experimental control.

Within each run, the three experimental conditions were executed sequentially. GPU memory was cleared between experiments via garbage collection and explicit CUDA cache emptying to prevent contamination. Each experiment followed an identical protocol: data splitting, tokenisation, class weight computation, custom Trainer instantiation, training, checkpoint selection by validation macro F1, and test set evaluation exclusively through `Trainer.predict()` followed by `sklearn.metrics.classification_report`. All reported test metrics derive from this single consistent evaluation path.

Figure 3 presents the training set class distribution on a log scale, confirming the long-tail structure.

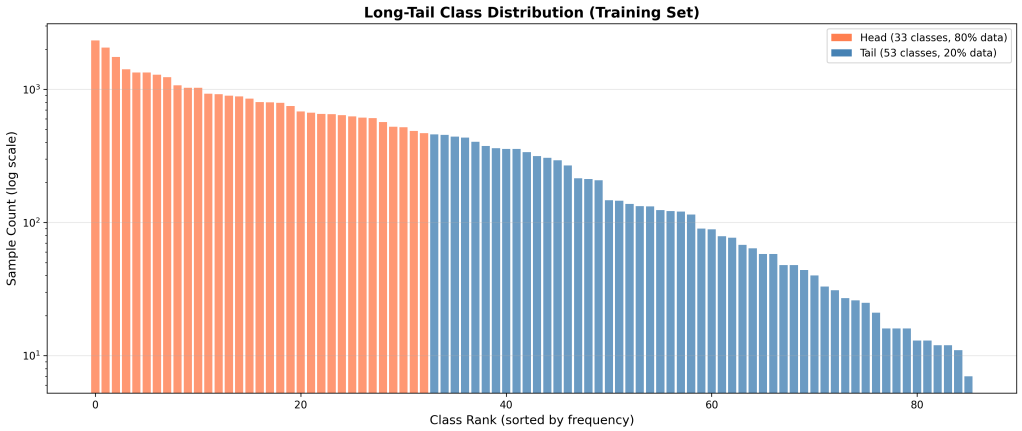


Figure 3. Long-tail class distribution of the training set on a log scale. Coral bars indicate head classes covering 80% of training data; steel-blue bars indicate tail classes.

The distribution in Figure 3 illustrates why standard fine-tuning fails on this dataset. The descent from the most frequent to the rarest agency spans more than two orders of magnitude, creating a training signal dominated by a small number of majority agencies. The Badan Perencanaan Pembangunan Daerah alone contributes 2,330 training samples, whereas seven agencies retain fewer than 10 samples each. The gradient contribution of tail classes under uniform cross-entropy is negligible relative to that of head classes.

5. Results

5.1. Overall Performance

Table 4 reports overall test set metrics averaged across three random seeds, expressed as mean ± standard deviation.

Table 4. Overall test set performance across three methods (mean ± std, three random seeds)

Method	Macro F1	Weighted F1	Accuracy	Macro Prec.	Macro Rec.
Baseline	0.3833 ± 0.0021	0.6212 ± 0.0014	0.6377 ± 0.0006	0.4189 ± 0.0105	0.3764 ± 0.0004
Class weighting	0.4441 ± 0.0029	0.5795 ± 0.0022	0.5666 ± 0.0038	0.4561 ± 0.0086	0.4939 ± 0.0053
Focal Loss	0.4241 ± 0.0055	0.5416 ± 0.0109	0.5207 ± 0.0098	0.4330 ± 0.0057	0.5007 ± 0.0043

Table 4 reveals three patterns. First, class weighting achieves the highest macro F1 at 0.4441 ± 0.0029, representing a 15.8% gain over the baseline (0.3833 ± 0.0021). Focal loss yields a 10.6% gain at 0.4241 ± 0.0055. Seed-level 95% confidence intervals for macro F1, computed with the t distribution on two degrees of freedom, are $[0.378, 0.388]$ for the baseline, $[0.437, 0.451]$ for class weighting, and $[0.410, 0.438]$ for focal loss; the intervals are wide relative to the seed standard deviations because only three seeds are available, and they do not overlap between the baseline and either cost-sensitive method. The finding that class weighting outperforms focal loss on overall macro F1 is noteworthy because focal loss was specifically designed to address class imbalance through difficulty weighting. Second, weighted F1 and accuracy decline under both cost-sensitive methods relative to the baseline. This divergence between macro and weighted metrics signals redistribution of predictive capacity from majority to minority classes, which is the intended behaviour. Third, variance across seeds is low for all methods, with coefficient of variation below 1.4% for all macro F1 estimates, indicating stable training dynamics under the adopted configuration.

The Friedman test comparing macro F1 across three methods and three seeds returns $\chi^2(2) = 6.00$, $p = 0.050$. With $k = 3$ methods and $n = 3$ seeds this is a boundary result: 6.0 is exactly the critical value at $\alpha = 0.05$, attained only because the method ranking is perfectly stable, with class weighting first, focal loss second, and the baseline third in every seed. The test is therefore read as evidence of consistent directional ordering under minimal statistical power rather than as conclusive proof of a method effect. Normality of the paired per-class F1 differences (class weighting minus baseline, pooled over all 86 classes) was assessed via the Shapiro-Wilk test prior to selecting between a parametric and non-parametric comparison. The test rejected normality ($W = 0.8541$, $p < 0.001$). Inspection of the difference distribution attributes this to heavy left-skew: many tail classes attain F1 of zero under the baseline, so the corresponding differences cluster at large positive values while head-class differences cluster near zero or slightly negative, producing a non-symmetric distribution. This result confirms the non-parametric Wilcoxon test, rather than a paired t-test, as the appropriate choice for the comparisons that follow. Table 5 summarises the per-seed statistical findings for the Wilcoxon comparison between baseline and class weighting.

Table 5. Per-seed Wilcoxon signed-rank test results: baseline versus class weighting, pooled over all 86 classes. Normality was rejected by Shapiro-Wilk ($W = 0.8541$, $p < 0.001$), justifying the non-parametric test.

Seed	W	p-value	Mean improvement	Cohen's d	Significant
42	1089.0	0.0753	0.0644	0.298	No
123	1123.0	0.0771	0.0594	0.309	No
2024	1160.0	0.1475	0.0584	0.309	No
Mean	1124.0	0.100	0.0607	0.305	–

Per-seed Wilcoxon tests comparing per-class F1 distributions yield p values of 0.0753, 0.0771, and 0.1475 across the three seeds, none reaching significance at the standard $\alpha = 0.05$ threshold. Mean Cohen's d stands at 0.305 ± 0.006, a small effect by conventional criteria. The pooled test, however, aggregates two populations that

the head/tail hypothesis predicts will move in opposite directions, so its non-significance does not indicate absence of an effect. Table 6 reports the segment-level decomposition.

Table 6. Segment-level Wilcoxon tests and bootstrap confidence intervals for per-class F1 differences (class weighting minus baseline), per seed

Seed	Segment	n	Mean diff.	Wilcoxon p	95% bootstrap CI
42	Tail	53	$+0.1460$	< 0.001	$[+0.089, +0.209]$
42	Head	33	-0.0667	< 0.001	$[-0.099, -0.036]$
123	Tail	53	$+0.1375$	< 0.001	$[+0.086, +0.193]$
123	Head	33	-0.0658	< 0.001	$[-0.096, -0.038]$
2024	Tail	53	$+0.1299$	< 0.001	$[+0.078, +0.186]$
2024	Head	33	-0.0564	< 0.001	$[-0.085, -0.031]$

Table 6 resolves the apparent tension between the non-significant pooled tests and the visibly large tail improvements. Within the tail segment, the F1 improvement is significant in every seed (two-sided p between 6.4×10^{-5} and 2.3×10^{-4}), and within the head segment the decline is likewise significant in every seed. All six bootstrap confidence intervals exclude zero. The pooled Wilcoxon statistic is diluted because 53 positive tail differences and 33 negative head differences partially cancel in the signed ranking. The correct statistical summary of the comparison is therefore not that the improvement fails to reach significance, but that class weighting produces two opposing, individually significant, seed-stable effects whose net direction favours the tail, consistent with the redistribution mechanism the method is designed to induce.

5.2. Head and Tail Class Analysis

Table 7 presents macro F1 separately for head and tail class segments, averaged across three seeds.

Table 7. Head and tail macro F1 by method (mean $\pm$ std, three seeds)

Method	Head F1	Tail F1	Head Change	Tail Change
Baseline	0.6700 ± 0.0006	0.2048 ± 0.0030	–	–
Class weighting	0.6070 ± 0.0057	0.3426 ± 0.0083	-9.4%	$+67.3\%$
Focal Loss	0.5669 ± 0.0123	0.3352 ± 0.0059	-15.4%	$+63.7\%$

The head/tail breakdown in Table 7 clarifies the mechanism underlying the overall macro F1 gains. The baseline achieves strong head performance at 0.6700 ± 0.0006 while producing tail macro F1 of only 0.2048 ± 0.0030, a level insufficient for operational routing of minority agency complaints. Class weighting raises tail F1 by 67.3% to 0.3426 ± 0.0083 at the cost of a 9.4% head decline. Focal loss achieves a 63.7% tail gain to 0.3352 ± 0.0059 but incurs a more severe 15.4% head reduction. Both head and tail results for class weighting are directionally consistent across all three seeds, and class weighting exceeds focal loss on both overall and tail macro F1 in every seed condition.

Figure 4 visualises these comparisons. In absolute macro F1 points, class weighting exchanges a mean head decline of 0.063 for a mean tail gain of 0.138, a gain-to-loss ratio of 2.19:1, whereas focal loss exchanges a 0.103 head decline for a 0.130 tail gain, a ratio of 1.26:1. The percentage changes in Table 7 are relative to segment baselines of very different magnitude and therefore overstate the exchange rate; the absolute ratios are the appropriate basis for comparing efficiency. This difference in efficiency favours class weighting for production systems where both head and tail agencies serve genuine citizen needs. The higher head variance for focal loss (± 0.0123) compared to class weighting (± 0.0057) also indicates less stable head performance, an additional operational concern.

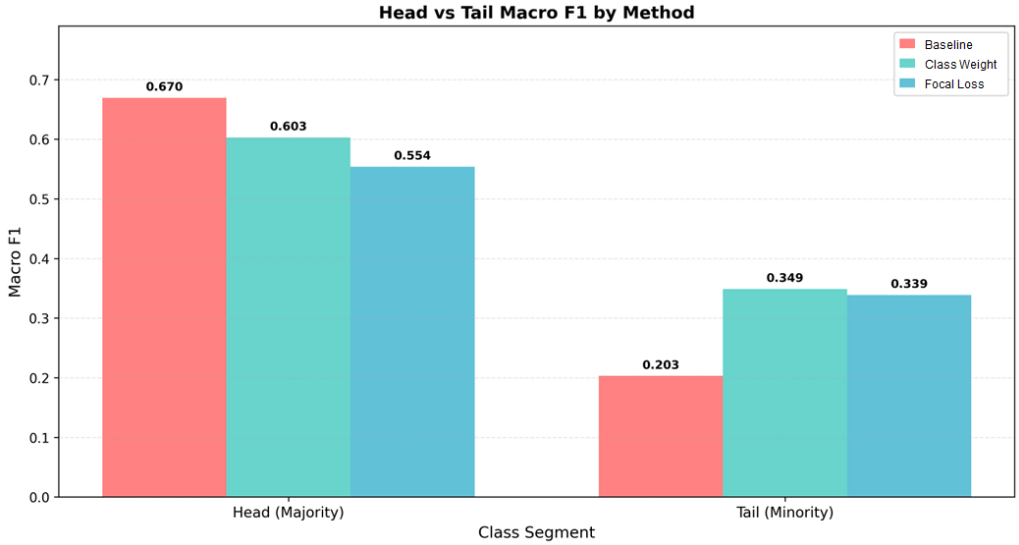


Figure 4. Macro F1 for head and tail class segments across three methods (seed 42). Bar values reflect the single-seed run; multi-seed means and standard deviations are reported in Table 7.

5.3. Per-Class Improvement Analysis

Figure 5 plots per-class F1 improvement (class weighting minus baseline) against baseline F1 for each of the 86 agencies.

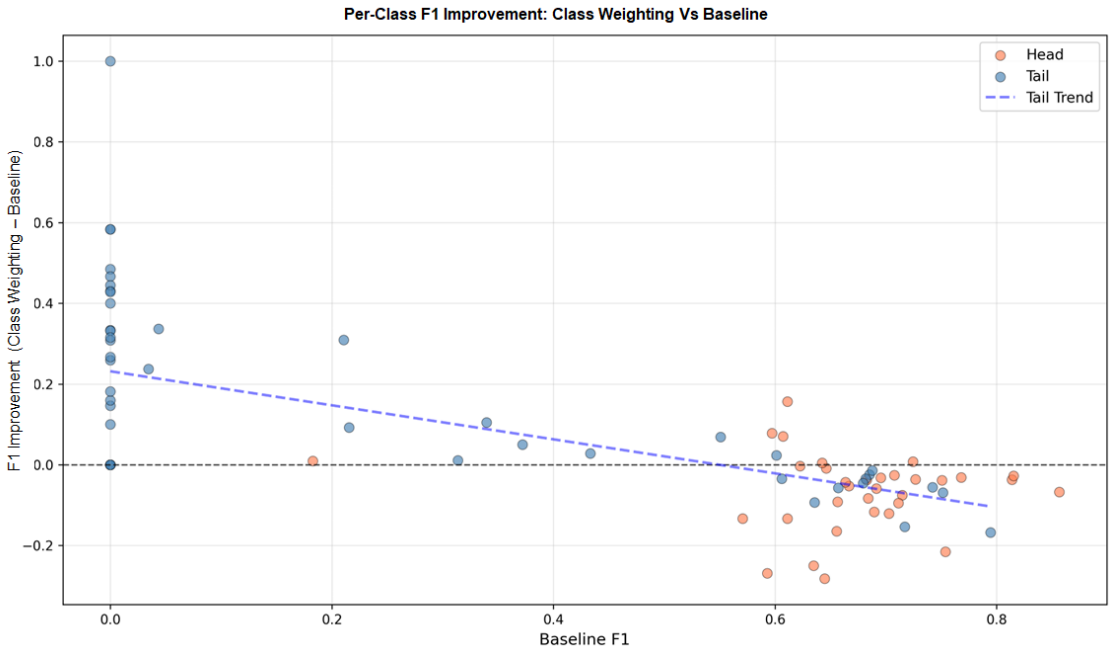


Figure 5. Per-class F1 improvement (class weighting minus baseline) versus baseline F1. Coral points indicate head classes; steel-blue points indicate tail classes. The dashed line is the linear trend for tail classes.

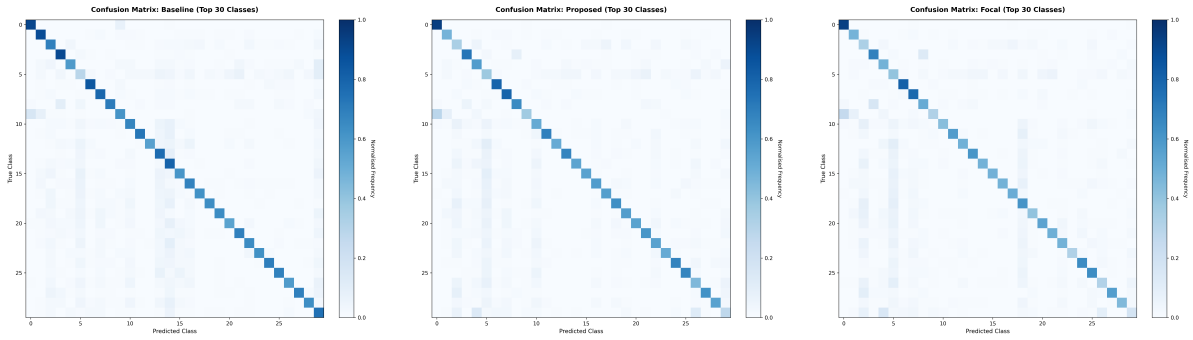
The scatter plot reveals a structured pattern. Across three seeds, an average of 33.7 classes improved (39.1% of all 86), while an average of 39.7 classes degraded and 12.7 remained unchanged. A majority of individual classes therefore decline or stay unchanged, and the aggregate macro F1 gain is driven by a minority of classes, predominantly in the tail, whose improvements are large. This asymmetry is a direct consequence of the redistribution mechanism and is reported here as a headline result rather than a footnote: whether the trade is acceptable depends on the operational weight attached to minority-agency coverage. Among the 53 tail classes, an average of 28.0 classes improved per seed. Averaged over all 53 tail classes, the mean F1 change is $+0.1378 \pm 0.0081$ (13.78 percentage points); averaged over all 33 head classes, the mean change is -0.063 ± 0.006 F1 points.

Spearman rank correlation between class training frequency and F1 improvement is consistently negative across seeds, with values of -0.588, -0.600, and -0.568 (mean $\rho = -0.585 \pm 0.016$, $p < 0.001$ in all three cases). This indicates that lower-frequency agencies benefit preferentially from inverse-frequency weighting, while the moderate magnitude of the correlation, corresponding to roughly one third of rank variance, shows that frequency is not the sole determinant of improvement; residual factors plausibly include semantic overlap between agency jurisdictions, label ambiguity, and per-class annotation quality. The negative correlation is visible in Figure 5 as a downward slope from left to right: classes with low baseline F1 (predominantly tail agencies with sparse training data) show the largest positive improvements, while classes with high baseline F1 (predominantly head agencies) tend to decline modestly.

Per-class metrics for all 86 agencies under each experimental condition are provided in Appendix B (baseline), Appendix C (class weighting), and Appendix D (focal loss) for full transparency. Representative cases from seed 42 illustrate the three regimes. Among tail agencies, BPJS Ketenagakerjaan and the provincial Animal Husbandry and Health agency both move from zero baseline F1 to 0.583 under class weighting. Mid-frequency agencies such as the Grobogan regency government are essentially unchanged (0.622 to 0.620). Among head agencies, the Semarang city government declines from 0.645 to 0.363, the largest single head-class concession, consistent with the aggregate pattern above.

5.4. Confusion Matrix Analysis

Figure 6 displays normalised confusion matrices for the top 30 most frequent agencies under all three experimental conditions.



(a) Baseline (b) Proposed (c) Focal Loss

Figure 6. Normalised confusion matrices for the top 30 agencies by test frequency, showing baseline (left), class weighting (centre, rendered as "Proposed" in the panel title), and focal loss (right).

In the baseline matrix, off-diagonal errors concentrate toward high-frequency head predictions. Cost-sensitive training strengthens minority class diagonals: the class weighting method reduces systematic misrouting of tail agencies toward majority ones. Certain confusion pairs persist across all three conditions, particularly between agencies with overlapping jurisdictions, such as Public Works Highways and Public Works Water Resources, reflecting genuine semantic similarity rather than model failure. Focal loss produces slightly stronger individual

tail diagonals in some agencies but at the cost of increased inter-class confusion among head agencies, consistent with its more aggressive down-weighting of easy examples.

The visual matrices in Figure 6 are restricted to the top 30 agencies and their differences are subtle to the eye; a quantitative summary restricted to zero-recall tail agencies, meaning agencies for which the model fails to correctly route a single test complaint, is therefore reported directly from the per-class metrics. Under the baseline, a mean of 33.0 of the 53 tail agencies (62.3%) have zero recall across the three seeds (range 32 to 34). Class weighting reduces this to a mean of 12.7 agencies (23.9%, range 12 to 13), and focal loss reduces it further to a mean of 9.7 agencies (18.3%, range 9 to 10). Focal loss therefore recovers more previously unrouted tail agencies than class weighting despite its lower mean tail F1, indicating that focal loss redistributes predictive capacity across a broader set of tail agencies at lower average confidence per agency, whereas class weighting concentrates larger gains on a somewhat narrower set. A full misrouting flow analysis, quantifying the fraction of tail-class errors specifically absorbed by head-class predictions, would require the complete confusion matrix over all 86 classes rather than the top-30 visual subset and is identified as a direct extension for future work.

5.5. *Training Dynamics*

Figure 7 plots training loss and validation macro F1 across epochs for all three methods.

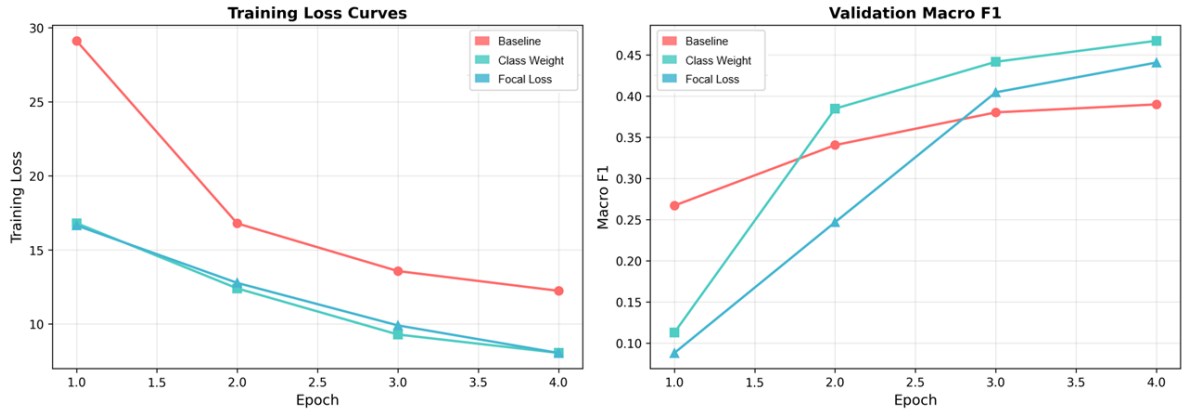


Figure 7. Training loss (left) and validation macro F1 (right) across four epochs for three methods (seed 42 re-run).

All three methods converge within four epochs. Cost-sensitive methods produce higher training loss than the baseline because they penalise minority-class errors more heavily, reducing the effective contribution of easy majority-class examples to gradient updates. Validation macro F1 for the baseline peaks early and then stabilises, whereas both cost-sensitive methods show sustained improvement across epochs. The separation between training and test performance narrows for cost-sensitive configurations, indicating reduced majority-class overfitting. Focal loss exhibits greater epoch-to-epoch variability in training loss, consistent with the dynamic re-weighting of individual examples, while class weighting produces a smoother trajectory.

5.6. *Practical Significance*

Beyond F1 improvements, the operational implications are quantifiable. The appropriate metric for translating model performance into a routing-accuracy estimate is recall weighted by test-set support within the tail segment, which equals the fraction of tail complaints assigned to their correct agency, rather than macro F1, which weights every class equally regardless of complaint volume and is not itself a routing-accuracy rate. Averaged across three seeds, the support-weighted tail recall rises from 38.53% under the baseline to 55.54% under class weighting, a gain of 17.01 percentage points, and to 54.94% under focal loss, a gain of 16.40 points. Tail agencies account for a mean of 20.44% of test-set complaint volume. The Jateng Ngopeni Nglakoni platform receives approximately 1,000 daily complaints, totalling roughly 30,000 monthly; applying the observed tail volume share

yields approximately 6,131 tail complaints per month. At a 17.01 percentage point improvement in weighted tail recall, this corresponds to approximately 1,043 additional correctly routed tail complaints per month under class weighting, and approximately 1,006 under focal loss. These estimates assume stationary complaint distribution and agency jurisdiction, both of which are subject to temporal drift in production, and assume that the observed test-set tail volume share and weighted recall generalise to future monthly traffic.

The Pareto-based evaluation framework itself constitutes a transferable methodological contribution. Standard accuracy metrics obscure imbalanced classifier performance by conflating majority and minority class outcomes. The proposed segmentation makes trade-offs explicit and reproducible, applicable to any domain with majority/minority class structure, including customer support routing, medical diagnosis coding, and legal document classification.

6. Discussion

6.1. Interpretation of Key Findings

Cost-sensitive loss modification effectively transfers predictive capacity from majority to minority agencies in extreme long-tail text classification. The baseline, despite achieving aggregate accuracy of 0.638, produces zero F1 on 33 of the 53 tail agencies (mean across seeds, 62.3%), leaving the majority of minority agencies with no correctly routed test complaints. Class weighting raises tail macro F1 by 67.3% to 0.343, while focal loss achieves a 63.7% gain. Both methods substantially exceed the baseline on the metric most relevant to service equity.

The finding that class weighting outperforms focal loss on overall macro F1, tail macro F1, and overall accuracy across all three seeds warrants interpretation. Focal loss was specifically designed to address training imbalance by concentrating gradient updates on hard examples. Under standard conditions, this produces superior results. In the present 332.86:1 setting, however, the gradient modulation from $(1 - p_t)^\gamma$ interacts with the extreme class weight ratios in a way that destabilises head class performance without producing commensurate tail gains. The 15.4% head decline under focal loss versus 9.4% for class weighting, at lower tail gains, suggests over-suppression of head-class gradients. This interpretation is qualified by the zero-recall analysis in Section 5: focal loss recovers more previously unrouted tail agencies (mean 9.7 of 53 remain at zero recall, versus 12.7 for class weighting), so the lower mean tail F1 under focal loss reflects lower average confidence across a broader set of recovered agencies rather than narrower coverage. This finding echoes observations by Zhang et al. [14] that methods calibrated for moderate imbalance may not transfer directly to extreme long-tail conditions.

The consistently negative Spearman correlation (mean $\rho = -0.585$) between class frequency and F1 improvement establishes that the mechanism operates as intended: agencies receiving fewer training samples benefit most from the inverse-frequency penalty. This is not a universal finding for all cost-sensitive methods; over-aggressive weighting can harm rare classes by assigning disproportionate gradient weight to noisy labels. The moderate Cohen's d of 0.305 indicates that the benefit is real but bounded by irreducible label-ambiguity and vocabulary overlap across agencies.

The Friedman test result ($p = 0.050$) is, as noted in Section 5, a boundary case driven by perfectly consistent method ranking across only three seeds, and is not treated here as strong evidence in its own right. The segment-level Wilcoxon tests and bootstrap confidence intervals in Table 6 provide the more informative statistical picture: the tail improvement and head decline are each individually significant in every seed, and the overall macro F1 gain is the net effect of two robust, opposing, within-segment changes rather than a marginal aggregate signal.

6.2. Comparison with Related Work

Table 8 positions the present results relative to prior studies on imbalanced classification for government or citizen-facing complaint data. Because the compared studies report different outcome metrics under different task definitions, the table lists each study's own reported improvement measure rather than a common tail-gain figure, and records separately whether a head/tail decomposition was performed.

Table 8 shows that this dataset carries the most extreme imbalance ratio among the compared studies, and that none of the three prior studies reports a head/tail decomposition or a tail-specific improvement figure; each

Table 8. Qualitative comparison with related studies on imbalanced government or citizen complaint classification. Reported improvement figures are each study's own outcome metric and are not directly commensurable across rows; see running text for definitions.

Study	Size	Classes	Ratio	Reported improvement	Head/tail analysis
Zahro et al. [4]	53,774	18	n.r.	Macro F1 0.606→0.625	None
Kim & Hong [37]	16,000	21→14	n.r.	Accuracy ≈90% (final stage)	None
Tang et al. [38]	3,438	16	≈130:1	Avg. accuracy 0.907→0.918	None
This study (class weighting)	56,068	86	332.86:1	Tail macro F1 0.205→0.343 (+67.3%)	Pareto-based
This study (focal loss)	56,068	86	332.86:1	Tail macro F1 0.205→0.335 (+63.7%)	Pareto-based

reports only an aggregate metric over all classes. This absence is itself informative; it means that whether these prior methods redistribute performance toward minority classes, or simply preserve majority-class accuracy while making limited headway on minority classes, cannot be determined from the published results, which is precisely the diagnostic gap that the Pareto-based convention adopted here is designed to close. Direct numerical comparison of tail-specific gains against these studies is therefore not possible, and no such comparison is claimed. Kim and Hong [37] and Tang et al. [38] address substantially fewer categories with re-sampling and augmentation rather than loss modification; Tang et al. operate under an imbalance ratio, reconstructed from their published category counts, of approximately 130:1, and Kim and Hong do not report a numeric ratio. Zahro et al. [4], the study closest in method and setting, apply focal loss alone without a competing cost-sensitive variant or head/tail breakdown. The present results indicate that loss-function approaches remain applicable at 332.86:1 imbalance without data manipulation, preserving the original corpus distribution during training, and that a head/tail decomposition changes the empirical picture materially: without it, the present results would be reported only as the 15.8% overall macro F1 gain in Table 4, obscuring the much larger and separately significant tail-specific effect.

6.3. Deployment Implications

Method selection in production should depend on the operational priority structure of the routing system. Class weighting provides the more balanced trade-off, with higher overall macro F1 and an absolute gain-to-loss ratio of 2.19:1 (Section 5), making it the preferable default for systems serving a diverse citizen population across all 86 agencies. Focal loss remains relevant in settings where tail-class coverage takes absolute priority and a head macro F1 near 0.567 (Table 7) is operationally acceptable, such as dedicated complaint pathways for specialised tail agencies, for example the provincial legal affairs bureau (Biro Hukum) or the provincial narcotics agency (Badan Narkotika Nasional Provinsi Jawa Tengah), both of which appear as tail agencies in Appendix A. Quarterly retraining on fresh complaint data would address temporal distribution drift and should be complemented by per-agency routing accuracy monitoring to detect emerging misrouting patterns.

7. Limitations

Single-dataset scope limits generalisability. The Central Java Province data reflect the administrative structure, language distribution, and service priorities of one provincial government. Dialectal variation exists across Indonesia's 34 provinces, with other regions using Sundanese, Balinese, and distinct local languages alongside Indonesian. Complaint topics and jurisdictional boundaries also differ by region. Multi-province validation studies would be required before asserting broad applicability of the cost-sensitive framework.

Temporal structure was not preserved in data splitting. The random stratified split ignores chronological ordering, which is appropriate for establishing a reproducible experimental benchmark but does not reflect the temporal sequence in which complaints arrive in production. Government priorities shift, agency jurisdictions merge or divide, and seasonal patterns affect complaint distribution. Deployment requires temporal-split evaluation to assess degradation over time and periodic retraining to maintain model currency.

The pooled per-seed Wilcoxon tests comparing baseline and class weighting across all 86 classes did not reach significance at $\alpha = 0.05$ in any individual seed (p values 0.075, 0.077, 0.147), because the pooled test combines a significant tail improvement with a significant head decline that partially cancel in the signed ranking (Table 6). The segment-level tests and bootstrap confidence intervals substantially strengthen the statistical picture, but they remain conditional on the correctness of the Pareto-based head/tail partition and on three seeds. The Friedman test result of $p = 0.050$ is a boundary case that should be interpreted cautiously. Only three random seeds were used, which is sufficient to establish directional consistency but limits the statistical power of any cross-seed test, including the Friedman test; a boundary result obtained from three seeds could plausibly shift with additional runs, and this should be verified before the finding is treated as robust to seed selection. Multi-site validation with independent datasets would substantially strengthen generalisation claims.

The focal loss condition reuses the inverse-frequency alpha weights from the class weighting method, so the reported focal loss results measure the incremental effect of difficulty-based modulation on top of frequency weighting rather than the effect of the unweighted focal loss formulation evaluated in the original proposal. An unweighted focal condition, and gradient norm diagnostics over training to verify that the large weight range (0.196 to 65.388) does not destabilise optimisation, are direct extensions that the present single-GPU-hour budget per seed did not accommodate.

Hyperparameter sensitivity remains uncharacterised. The learning rate (2×10^{-5}) and focal γ (2.0) were adopted from prior literature rather than optimised for this dataset. Systematic tuning through grid search or Bayesian optimisation could yield further improvements. An adaptive, imbalance-ratio-conditioned loss selection strategy, such as that proposed by Winarno et al. [17], offers a concrete starting framework for this tuning problem. The computational cost of approximately 4.4 GPU-hours per seed renders exhaustive search infeasible within the present resource envelope.

Class filtering excluded 24 agencies with fewer than 10 training samples, affecting 21.8% of original agencies, so the evaluated model covers 86 of the 110 registered agencies and does not address the most extreme part of the long tail. The evaluated model therefore does not cover the full operational agency roster. Ultra-rare agencies with one to nine training samples represent a distinct challenge that standard fine-tuning under the present protocol cannot address, since stratified splitting requires a minimum sample count per partition. Few-shot or prototypical-network approaches, which learn a class representation from a handful of labelled examples rather than from gradient-based fine-tuning on the full class set, are identified as the necessary next step for these agencies and constitute a concrete direction for future work rather than an incidental remark.

The multi-seed validation presented here used three seeds, which suffices to establish directional consistency and stable variance estimates but provides limited statistical power for formal hypothesis testing across seeds. Expanding to five or more seeds, combined with cross-validation on temporal data partitions, would provide more robust estimates of the true performance distribution.

8. Conclusion

This study compared two cost-sensitive loss modification methods for extreme long-tail intent classification in Indonesian government complaint routing, a setting where standard transformer fine-tuning leaves most minority agencies without correct routings due to a 332.86:1 class imbalance ratio. Experiments were conducted across 86 government agencies, 56,068 code-mixed complaints, and three independent random seeds, providing stable, reproducible estimates of method performance.

Class weighting achieved overall macro F1 of 0.4441 ± 0.0029, a 15.8% improvement over the baseline (0.3833 ± 0.0021), with tail macro F1 rising by 67.3% from 0.2048 to 0.3426 and a corresponding 9.4% head decline. Focal loss produced a 10.6% overall macro F1 gain and a 63.7% tail gain, but incurred a 15.4% head reduction. Class weighting outperformed focal loss on overall macro F1, tail macro F1, and overall accuracy across all three seeds, a finding that runs counter to the expectation that focal loss, specifically designed for imbalance, would dominate under extreme long-tail conditions. Support-weighted tail recall, the metric most directly tied

to routing accuracy, rose from 38.53% to 55.54% under class weighting, corresponding to an estimated 1,043 additional correctly routed tail complaints per month on the deployed platform under stated volume assumptions.

Statistical evidence for the improvement rests on two complementary results. The Friedman test across three methods and three seeds returns a boundary result ($\chi^2(2) = 6.00$, $p = 0.050$) driven by perfectly consistent method ranking, which is read as directional rather than conclusive given the limited power of three seeds. Segment-level Wilcoxon tests and bootstrap confidence intervals provide stronger evidence: the tail improvement and head decline are each individually significant in every seed ($p < 0.001$), and all six bootstrap confidence intervals over classes exclude zero, showing that the non-significant pooled test reflects two robust opposing effects rather than the absence of an effect. Spearman rank correlation between class training frequency and F1 improvement is consistently negative across seeds (mean $\rho = -0.585$, $p < 0.001$), confirming that inverse-frequency weighting preferentially benefits agencies with the fewest training samples, though the moderate correlation magnitude indicates that frequency alone does not fully determine which classes benefit.

The Pareto-based evaluation convention adopted here provides a principled, threshold-free decomposition of classifier performance into head and tail contributions. Unlike fixed-threshold definitions of minority classes, Pareto segmentation adapts to the empirical data distribution and makes head/tail trade-offs explicit and reproducible. This decomposition also surfaces a result that an aggregate metric alone would conceal: a majority of individual agencies (a mean of 39.7 of 86 per seed) see degraded or unchanged F1 under class weighting, and the overall gain is carried by large improvements concentrated in a minority of, predominantly tail, agencies. Whether this trade is acceptable is an operational judgement that depends on the relative weight assigned to majority and minority agency service quality, not a judgement this study makes on the deployer's behalf. This framework is applicable to any imbalanced classification domain, including customer support routing, medical diagnosis coding, and multi-label document classification, wherever accurate reporting of minority class performance is operationally relevant.

Future research directions include temporal-split evaluation to assess model degradation under distribution shift, multi-province validation to test geographic generalisability, few-shot learning approaches for agencies below the minimum frequency threshold, an unweighted focal loss ablation to isolate the effect of difficulty modulation from frequency weighting, gradient norm diagnostics to verify optimisation stability under the observed weight range, evaluation of performance uniformity across code-mixing patterns, and Bayesian hyperparameter optimisation for learning rate and focal γ under the extreme imbalance conditions characterised here.

Acknowledgement

The authors thank the Communication and Information Agency of Central Java Province (Diskominfo Jawa Tengah) for providing access to the Jateng Ngopeni Nglakoni operational complaint data. Computational experiments were conducted using Kaggle GPU resources.

This research was supported by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the BIMA Fundamental Research Grant (Penelitian Fundamental Reguler, Multiyear Program Year 2), under Master Contract No. 133/C3/DT.05.00/PL-MULTITAHUN LANJUTAN/2026 and Derivative Contracts No. 16/LL6/AL.04.03/PL-MULTITAHUN LANJUTAN/2026 and No. 131/F.09/UDN-09/III/2026

A. Government Agency List

Table 9 lists all 86 government agencies included in the experimental dataset, sorted by class identifier. The Segment column indicates Pareto classification (Head or Tail); the Support column reports the number of test set samples assigned to each agency. Agencies numbered 1 through 86 correspond to class identifiers 0 through 85 in the label encoding used throughout the pipeline.

Table 9. All 86 government agencies with Pareto segment and test set support

No.	Agency	Segment	Support
1	BADAN KEPEGAWAIAN DAERAH	Head	31
2	BADAN KESATUAN BANGSA DAN POLITIK	Tail	2
3	BADAN NARKOTIKA NASIONAL PROVINSI JAWA TENGAH	Tail	16
4	BADAN PENANGGULANGAN BENCANA DAERAH	Tail	3
5	BADAN PENGELOLA PENDAPATAN DAERAH	Head	182
6	BADAN PENGEMBANGAN SUMBER DAYA MANUSIA DAERAH	Head	108
7	BADAN PERENCANAAN PEMBANGUNAN, PENELITIAN DAN PENGEMBANGAN DAERAH	Head	2330
8	BIRO ADMINISTRASI PEMBANGUNAN DAERAH	Tail	7
9	BIRO HUKUM	Tail	17
10	BIRO PEREKONOMIAN	Tail	12
11	BIRO UMUM	Tail	14
12	BPJS Kesehatan	Tail	72
13	BPJS Ketenagakerjaan Kanwil Jateng dan DIY	Tail	10
14	BPTD Kelas I Jawa Tengah	Tail	26
15	Balai Besar Pelaksanaan Jalan Nasional Jawa Tengah-DIY	Tail	93
16	Balai Besar Wilayah Sungai Pemali Juana	Tail	15
17	DINAS ENERGI DAN SUMBER DAYA MINERAL	Head	221
18	DINAS KEARSIPAN DAN PERPUSTAKAAN	Tail	3
19	DINAS KELAUTAN DAN PERIKANAN	Tail	9
20	DINAS KEPEMUDAAN, OLAHRAGA DAN PARIWISATA	Tail	17
21	DINAS KESEHATAN	Tail	67
22	DINAS KOMUNIKASI DAN INFORMATIKA	Tail	19
23	DINAS KOPERASI DAN UMKM	Tail	28
24	DINAS LINGKUNGAN HIDUP DAN KEHUTANAN	Tail	29
25	DINAS PEKERJAAN UMUM BINA MARGA CIPTA KARYA	Head	276
26	DINAS PEKERJAAN UMUM SUMBER DAYA AIR DAN PENATAAN RUANG	Tail	77
27	DINAS PEMBERDAYAAN MASYARAKAT, DESA, KEPENDUDUKAN DAN PENCATATAN SIPIL	Tail	57
28	DINAS PEMBERDAYAAN PEREMPUAN, PERLINDUNGAN ANAK, PP DAN KB	Tail	10
29	DINAS PENANAMAN MODAL DAN PELAYANAN TERPADU SATU PINTU	Tail	6
30	DINAS PENDIDIKAN DAN KEBUDAYAAN	Head	441
31	DINAS PERHUBUNGAN	Head	172
32	DINAS PERINDUSTRIAN DAN PERDAGANGAN	Tail	32
33	DINAS PERTANIAN DAN PERKEBUNAN	Tail	28
34	DINAS PERUMAHAN RAKYAT DAN KAWASAN PERMUKIMAN	Tail	45
35	DINAS PETERNAKAN DAN KESEHATAN HEWAN	Tail	9
36	DINAS SOSIAL	Head	122
37	DINAS TENAGA KERJA DAN TRANSMIGRASI	Head	304
38	INSPEKTORAT	Tail	7
39	Kanwil BPN Provinsi Jawa Tengah	Head	146
40	Kanwil Kemenag Provinsi Jawa Tengah	Head	112
41	Kepolisian Daerah Jawa Tengah	Head	287
42	Komisi Pemilihan Umum Provinsi Jawa Tengah	Tail	3
43	PLN Distribusi Jateng DIY	Tail	27
44	Pemda Kabupaten Banjarnegara	Tail	98
45	Pemda Kabupaten Banyumas	Head	230
46	Pemda Kabupaten Batang	Head	111
47	Pemda Kabupaten Blora	Head	132
48	Pemda Kabupaten Boyolali	Head	130
49	Pemda Kabupaten Brebes	Head	287
50	Pemda Kabupaten Cilacap	Head	375
51	Pemda Kabupaten Demak	Head	264

Table 9 (continued)

No.	Agency	Segment	Support
52	Pemda Kabupaten Grobogan	Head	220
53	Pemda Kabupaten Jepara	Head	138
54	Pemda Kabupaten Karanganyar	Tail	98
55	Pemda Kabupaten Kebumen	Head	139
56	Pemda Kabupaten Kendal	Head	170
57	Pemda Kabupaten Klaten	Head	160
58	Pemda Kabupaten Kudus	Tail	94
59	Pemda Kabupaten Magelang	Head	199
60	Pemda Kabupaten Pati	Head	197
61	Pemda Kabupaten Pekalongan	Head	144
62	Pemda Kabupaten Pemalang	Head	192
63	Pemda Kabupaten Purbalingga	Head	101
64	Pemda Kabupaten Purworejo	Head	134
65	Pemda Kabupaten Rembang	Tail	66
66	Pemda Kabupaten Semarang	Head	171
67	Pemda Kabupaten Sragen	Tail	77
68	Pemda Kabupaten Sukoharjo	Head	141
69	Pemda Kabupaten Tegal	Head	190
70	Pemda Kabupaten Temanggung	Tail	87
71	Pemda Kabupaten Wonogiri	Tail	63
72	Pemda Kabupaten Wonosobo	Tail	78
73	Pemda Kota Magelang	Tail	25
74	Pemda Kota Pekalongan	Head	105
75	Pemda Kota Salatiga	Tail	26
76	Pemda Kota Semarang	Head	499
77	Pemda Kota Surakarta	Tail	46
78	Pemda Kota Tegal	Tail	80
79	Pertamina UMPS IV Jateng dan DIY	Tail	13
80	Perum Perhutani Divisi Regional Jawa Tengah	Tail	4
81	RSUD MARGONO SOEKARJO BANYUMAS	Tail	4
82	RSUD MOEWARDI SURAKARTA	Tail	7
83	RSUD TUGUREJO SEMARANG	Tail	3
84	RSUP DR. KARIADI	Tail	3
85	SATPOL PP	Tail	5
86	SEKRETARIAT DPRD	Tail	1

B. Per-Class Metrics: Baseline

Table 10 reports precision, recall, F1, and test set support for all 86 agencies under the baseline CrossEntropyLoss configuration (seed 42).

Table 10. Per-class metrics for baseline method (seed 42)

No.	Agency	Prec.	Rec.	F1	Support
1	BADAN KEPEGAWAIAN DAERAH	0.448	0.419	0.433	31
2	BADAN KESATUAN BANGSA DAN POLITIK	0.000	0.000	0.000	2
3	BADAN NARKOTIKA NASIONAL PROVINSI JAWA TENGAH	0.000	0.000	0.000	16
4	BADAN PENANGGULANGAN BENCANA DAERAH	0.000	0.000	0.000	3
5	BADAN PENGELOLA PENDAPATAN DAERAH	0.735	0.912	0.814	182
6	BADAN PENGEMBANGAN SUMBER DAYA MANUSIA DAERAH	0.000	0.000	0.000	6

Table 10 (continued)

No.	Agency	Prec.	Rec.	F1	Support
7	BADAN PERENCANAAN PEMBANGUNAN DAN LITBANG DAERAH	0.597	0.804	0.685	46
8	BIRO ADMINISTRASI PEMBANGUNAN DAERAH	0.000	0.000	0.000	3
9	BIRO HUKUM	0.000	0.000	0.000	19
10	BIRO PEREKONOMIAN	0.000	0.000	0.000	12
11	BIRO UMUM	0.000	0.000	0.000	14
12	BPJS Kesehatan	0.538	0.681	0.601	72
13	BPJS Ketenagakerjaan Kanwil Jateng-DIY	0.000	0.000	0.000	10
14	BPTD Kelas I Jawa Tengah	0.000	0.000	0.000	26
15	Balai Besar Jalan Nasional Jawa Tengah-DIY	0.000	0.000	0.000	93
16	Balai Besar Wilayah Sungai Pemali Juana	0.000	0.000	0.000	15
17	DINAS ENERGI DAN SUMBER DAYA MINERAL	0.656	0.887	0.754	221
18	DINAS KEARSIPAN DAN PERPUSTAKAAN	0.000	0.000	0.000	3
19	DINAS KELAUTAN DAN PERIKANAN	0.000	0.000	0.000	9
20	DINAS KEPEMUDAAN, OLAHRAGA DAN PARIWISATA	1.000	0.118	0.211	17
21	DINAS KESEHATAN	0.387	0.358	0.372	67
22	DINAS KOMUNIKASI DAN INFORMATIKA	0.000	0.000	0.000	19
23	DINAS KOPERASI DAN UMKM	0.000	0.000	0.000	28
24	DINAS LINGKUNGAN HIDUP DAN KEHUTANAN	0.000	0.000	0.000	29
25	DINAS PEKERJAAN UMUM BINA MARGA CIPTA KARYA	0.521	0.688	0.593	276
26	DINAS PEKERJAAN UMUM SDA DAN PENATAAN RUANG	0.316	0.312	0.314	77
27	DINAS PEMBERDAYAAN MASYARAKAT, DESA DAN DUKCAPIL	1.000	0.018	0.034	57
28	DINAS PEMBERDAYAAN PEREMPUAN, PA, PP DAN KB	0.000	0.000	0.000	10
29	DINAS PENANAMAN MODAL DAN PTSP	0.000	0.000	0.000	6
30	DINAS PENDIDIKAN DAN KEBUDAYAAN	0.817	0.900	0.857	441
31	DINAS PERHUBUNGAN	0.555	0.587	0.571	172
32	DINAS PERINDUSTRIAN DAN PERDAGANGAN	0.429	0.281	0.340	32
33	DINAS PERTANIAN DAN PERKEBUNAN	0.189	0.250	0.215	28
34	DINAS PERUMAHAN RAKYAT DAN KAWASAN PERMUKIMAN	1.000	0.022	0.043	45
35	DINAS PETERNAKAN DAN KESEHATAN HEWAN	0.000	0.000	0.000	9
36	DINAS SOSIAL	0.134	0.287	0.183	122
37	DINAS TENAGA KERJA DAN TRANSMIGRASI	0.784	0.849	0.815	304
38	INSPEKTORAT	0.000	0.000	0.000	7
39	Kanwil BPN Provinsi Jawa Tengah	0.548	0.788	0.646	146
40	Kanwil Kemenag Provinsi Jawa Tengah	0.664	0.705	0.684	112
41	Kepolisian Daerah Jawa Tengah	0.709	0.610	0.655	287
42	KPU Provinsi Jawa Tengah	0.000	0.000	0.000	3
43	PLN Distribusi Jateng DIY	0.000	0.000	0.000	27
44	Pemda Kab. Banjarnegara	0.714	0.663	0.688	98
45	Pemda Kab. Banyumas	0.562	0.670	0.611	230
46	Pemda Kab. Batang	0.790	0.577	0.667	111
47	Pemda Kab. Blora	0.814	0.727	0.768	132
48	Pemda Kab. Boyolali	0.811	0.562	0.664	130
49	Pemda Kab. Brebes	0.508	0.767	0.611	287
50	Pemda Kab. Cilacap	0.476	0.803	0.597	375
51	Pemda Kab. Demak	0.608	0.606	0.607	264

Table 10 (continued)

No.	Agency	Prec.	Rec.	F1	Support
52	Pemda Kab. Grobogan	0.573	0.682	0.622	220
53	Pemda Kab. Jepara	0.754	0.623	0.683	138
54	Pemda Kab. Karanganyar	0.746	0.510	0.606	98
55	Pemda Kab. Kebumen	0.809	0.640	0.715	139
56	Pemda Kab. Kendal	0.769	0.647	0.703	170
57	Pemda Kab. Klaten	0.772	0.550	0.642	160
58	Pemda Kab. Kudus	0.782	0.723	0.751	94
59	Pemda Kab. Magelang	0.703	0.688	0.695	199
60	Pemda Kab. Pati	0.668	0.645	0.656	197
61	Pemda Kab. Pekalongan	0.826	0.625	0.711	144
62	Pemda Kab. Pemalang	0.812	0.698	0.751	192
63	Pemda Kab. Purbalingga	0.798	0.663	0.724	101
64	Pemda Kab. Purworejo	0.730	0.687	0.708	134
65	Pemda Kab. Rembang	0.829	0.515	0.636	66
66	Pemda Kab. Semarang	0.702	0.579	0.635	171
67	Pemda Kab. Sragen	0.818	0.584	0.682	77
68	Pemda Kab. Sukoharjo	0.770	0.688	0.727	141
69	Pemda Kab. Tegal	0.756	0.637	0.691	190
70	Pemda Kab. Temanggung	0.768	0.609	0.679	87
71	Pemda Kab. Wonogiri	0.754	0.730	0.742	63
72	Pemda Kab. Wonosobo	0.889	0.718	0.794	78
73	Pemda Kota Magelang	0.000	0.000	0.000	25
74	Pemda Kota Pekalongan	0.847	0.581	0.689	105
75	Pemda Kota Salatiga	0.704	0.731	0.717	26
76	Pemda Kota Semarang	0.566	0.749	0.645	499
77	Pemda Kota Surakarta	0.826	0.413	0.551	46
78	Pemda Kota Tegal	0.789	0.562	0.657	80
79	Pertamina UMPS IV Jateng-DIY	0.000	0.000	0.000	13
80	Perum Perhutani Divisi Regional Jawa Tengah	0.000	0.000	0.000	4
81	RSUD MARGONO SOEKARJO	0.000	0.000	0.000	4
82	RSUD MOEWARDI SURAKARTA	0.000	0.000	0.000	7
83	RSUD TUGUREJO SEMARANG	0.000	0.000	0.000	3
84	RSUP DR. KARIADI	0.000	0.000	0.000	3
85	SATPOL PP	0.000	0.000	0.000	5
86	SEKRETARIAT DPRD	0.000	0.000	0.000	1

C. Per-Class Metrics: Class Weighting Method

Table 11 reports per-class metrics for the inverse-frequency class weighting method (seed 42). Comparison with Appendix B reveals which agencies gained and which declined under cost-sensitive training.

Table 11. Per-class metrics for the class weighting method (seed 42)

No.	Agency	Prec.	Rec.	F1	Support
1	BADAN KEPEGAWAIAN DAERAH	0.350	0.677	0.462	31
2	BADAN KESATUAN BANGSA DAN POLITIK	0.000	0.000	0.000	2

Table 11 (continued)

No.	Agency	Prec.	Rec.	F1	Support
3	BADAN NARKOTIKA NASIONAL PROVINSI JAWA TENGAH	0.120	0.188	0.146	16
4	BADAN PENANGGULANGAN BENCANA DAERAH	0.000	0.000	0.000	3
5	BADAN PENGELOLA PENDAPATAN DAERAH	0.581	0.923	0.713	182
6	BADAN PENGEMBANGAN SUMBER DAYA MANUSIA DAERAH	0.385	0.833	0.526	6
7	BADAN PERENCANAAN PEMBANGUNAN DAN LITBANG DAERAH	0.529	0.587	0.556	46
8	BIRO ADMINISTRASI PEMBANGUNAN DAERAH	0.500	0.333	0.400	3
9	BIRO HUKUM	0.162	0.158	0.160	19
10	BIRO PEREKONOMIAN	0.091	0.083	0.087	12
11	BIRO UMUM	0.032	0.071	0.044	14
12	BPJS Kesehatan	0.537	0.597	0.566	72
13	BPJS Ketenagakerjaan Kanwil Jateng-DIY	0.480	0.600	0.533	10
14	BPTD Kelas I Jawa Tengah	0.148	0.154	0.151	26
15	Balai Besar Jalan Nasional Jawa Tengah-DIY	0.337	0.312	0.324	93
16	Balai Besar Wilayah Sungai Pemali Juana	0.176	0.200	0.187	15
17	DINAS ENERGI DAN SUMBER DAYA MINERAL	0.537	0.878	0.667	221
18	DINAS KEARSIPAN DAN PERPUSTAKAAN	1.000	0.333	0.500	3
19	DINAS KELAUTAN DAN PERIKANAN	0.600	0.333	0.429	9
20	DINAS KEPEMUDAAN, OLAHRAGA DAN PARIWISATA	0.438	0.412	0.424	17
21	DINAS KESEHATAN	0.365	0.552	0.440	67
22	DINAS KOMUNIKASI DAN INFORMATIKA	0.286	0.316	0.300	19
23	DINAS KOPERASI DAN UMKM	0.250	0.250	0.250	28
24	DINAS LINGKUNGAN HIDUP DAN KEHUTANAN	0.333	0.276	0.302	29
25	DINAS PEKERJAAN UMUM BINA MARGA CIPTA KARYA	0.434	0.576	0.495	276
26	DINAS PEKERJAAN UMUM SDA DAN PENATAAN RUANG	0.329	0.286	0.306	77
27	DINAS PEMBERDAYAAN MASYARAKAT, DESA DAN DUKCAPIL	0.393	0.193	0.258	57
28	DINAS PEMBERDAYAAN PEREMPUAN, PA, PP DAN KB	0.400	0.200	0.267	10
29	DINAS PENANAMAN MODAL DAN PTSP	0.333	0.167	0.222	6
30	DINAS PENDIDIKAN DAN KEBUDAYAAN	0.698	0.907	0.789	441
31	DINAS PERHUBUNGAN	0.470	0.564	0.513	172
32	DINAS PERINDUSTRIAN DAN PERDAGANGAN	0.310	0.281	0.295	32
33	DINAS PERTANIAN DAN PERKEBUNAN	0.340	0.607	0.435	28
34	DINAS PERUMAHAN RAKYAT DAN KAWASAN PERMUKIMAN	0.375	0.267	0.311	45
35	DINAS PETERNAKAN DAN KESEHATAN HEWAN	0.333	0.222	0.267	9
36	DINAS SOSIAL	0.209	0.361	0.265	122
37	DINAS TENAGA KERJA DAN TRANSMIGRASI	0.676	0.822	0.742	304
38	INSPEKTORAT	0.333	0.143	0.200	7
39	Kanwil BPN Provinsi Jawa Tengah	0.548	0.747	0.632	146
40	Kanwil Kemenag Provinsi Jawa Tengah	0.608	0.598	0.603	112
41	Kepolisian Daerah Jawa Tengah	0.617	0.608	0.612	287
42	KPU Provinsi Jawa Tengah	0.000	0.000	0.000	3
43	PLN Distribusi Jateng DIY	0.349	0.556	0.428	27
44	Pemda Kab. Banjarnegara	0.569	0.541	0.555	98
45	Pemda Kab. Banyumas	0.497	0.630	0.556	230
46	Pemda Kab. Batang	0.659	0.568	0.610	111

Table 11 (continued)

No.	Agency	Prec.	Rec.	F1	Support
47	Pemda Kab. Blora	0.691	0.720	0.706	132
48	Pemda Kab. Boyolali	0.712	0.554	0.623	130
49	Pemda Kab. Brebes	0.500	0.737	0.596	287
50	Pemda Kab. Cilacap	0.416	0.784	0.544	375
51	Pemda Kab. Demak	0.558	0.595	0.576	264
52	Pemda Kab. Grobogan	0.528	0.623	0.571	220
53	Pemda Kab. Jepara	0.670	0.609	0.638	138
54	Pemda Kab. Karanganyar	0.587	0.459	0.515	98
55	Pemda Kab. Kebumen	0.723	0.568	0.636	139
56	Pemda Kab. Kendal	0.694	0.612	0.650	170
57	Pemda Kab. Klaten	0.658	0.500	0.568	160
58	Pemda Kab. Kudus	0.625	0.638	0.632	94
59	Pemda Kab. Magelang	0.644	0.643	0.643	199
60	Pemda Kab. Pati	0.604	0.584	0.594	197
61	Pemda Kab. Pekalongan	0.759	0.618	0.681	144
62	Pemda Kab. Pemalang	0.722	0.724	0.723	192
63	Pemda Kab. Purbalingga	0.706	0.683	0.694	101
64	Pemda Kab. Purworejo	0.680	0.672	0.676	134
65	Pemda Kab. Rembang	0.716	0.500	0.589	66
66	Pemda Kab. Semarang	0.632	0.538	0.581	171
67	Pemda Kab. Sragen	0.708	0.571	0.633	77
68	Pemda Kab. Sukoharjo	0.723	0.674	0.698	141
69	Pemda Kab. Tegal	0.683	0.611	0.644	190
70	Pemda Kab. Temanggung	0.618	0.609	0.613	87
71	Pemda Kab. Wonogiri	0.727	0.730	0.729	63
72	Pemda Kab. Wonosobo	0.732	0.756	0.744	78
73	Pemda Kota Magelang	0.400	0.160	0.229	25
74	Pemda Kota Pekalongan	0.788	0.610	0.688	105
75	Pemda Kota Salatiga	0.679	0.731	0.704	26
76	Pemda Kota Semarang	0.501	0.737	0.597	499
77	Pemda Kota Surakarta	0.667	0.391	0.493	46
78	Pemda Kota Tegal	0.598	0.625	0.611	80
79	Pertamina UMPS IV Jateng-DIY	0.500	0.154	0.235	13
80	Perum Perhutani Divisi Regional Jawa Tengah	0.000	0.000	0.000	4
81	RSUD MARGONO SOEKARJO	0.333	0.250	0.286	4
82	RSUD MOEWARDI SURAKARTA	0.000	0.000	0.000	7
83	RSUD TUGUREJO SEMARANG	0.000	0.000	0.000	3
84	RSUP DR. KARIADI	0.000	0.000	0.000	3
85	SATPOL PP	0.000	0.000	0.000	5
86	SEKRETARIAT DPRD	0.000	0.000	0.000	1

D. Per-Class Metrics: Focal Loss

Table 12 reports per-class metrics for the focal loss configuration with $\gamma = 2.0$ and inverse-frequency alpha weighting (seed 42).

Table 12. Per-class metrics for focal loss ($\gamma = 2.0$, seed 42)

No.	Agency	Prec.	Rec.	F1	Support
1	BADAN KEPEGAWAIAN DAERAH	0.282	0.645	0.392	31
2	BADAN KESATUAN BANGSA DAN POLITIK	0.000	0.000	0.000	2
3	BADAN NARKOTIKA NASIONAL PROVINSI JAWA TENGAH	0.103	0.500	0.170	16
4	BADAN PENANGGULANGAN BENCANA DAERAH	0.000	0.000	0.000	3
5	BADAN PENGELOLA PENDAPATAN DAERAH	0.504	0.879	0.641	182
6	BADAN PENGEMBANGAN SUMBER DAYA MANUSIA DAERAH	0.188	0.500	0.273	6
7	BADAN PERENCANAAN PEMBANGUNAN DAN LITBANG DAERAH	0.440	0.478	0.458	46
8	BIRO ADMINISTRASI PEMBANGUNAN DAERAH	0.500	0.333	0.400	3
9	BIRO HUKUM	0.222	0.105	0.143	19
10	BIRO PEREKONOMIAN	0.250	0.083	0.125	12
11	BIRO UMUM	0.059	0.071	0.065	14
12	BPJS Kesehatan	0.476	0.556	0.513	72
13	BPJS Ketenagakerjaan Kanwil Jateng-DIY	0.500	0.600	0.545	10
14	BPTD Kelas I Jawa Tengah	0.120	0.115	0.118	26
15	Balai Besar Jalan Nasional Jawa Tengah-DIY	0.285	0.247	0.265	93
16	Balai Besar Wilayah Sungai Pemali Juana	0.222	0.267	0.242	15
17	DINAS ENERGI DAN SUMBER DAYA MINERAL	0.502	0.882	0.640	221
18	DINAS KEARSIPAN DAN PERPUSTAKAAN	0.000	0.000	0.000	3
19	DINAS KELAUTAN DAN PERIKANAN	0.400	0.222	0.286	9
20	DINAS KEPEMUDAAN, OLAHRAGA DAN PARIWISATA	0.389	0.412	0.400	17
21	DINAS KESEHATAN	0.356	0.433	0.391	67
22	DINAS KOMUNIKASI DAN INFORMATIKA	0.286	0.316	0.300	19
23	DINAS KOPERASI DAN UMKM	0.194	0.214	0.203	28
24	DINAS LINGKUNGAN HIDUP DAN KEHUTANAN	0.267	0.276	0.271	29
25	DINAS PEKERJAAN UMUM BINA MARGA CIPTA KARYA	0.384	0.554	0.454	276
26	DINAS PEKERJAAN UMUM SDA DAN PENATAAN RUANG	0.280	0.364	0.316	77
27	DINAS PEMBERDAYAAN MASYARAKAT, DESA DAN DUKCAPIL	0.339	0.193	0.246	57
28	DINAS PEMBERDAYAAN PEREMPUAN, PA, PP DAN KB	0.333	0.300	0.316	10
29	DINAS PENANAMAN MODAL DAN PTSP	0.143	0.167	0.154	6
30	DINAS PENDIDIKAN DAN KEBUDAYAAN	0.652	0.886	0.752	441
31	DINAS PERHUBUNGAN	0.404	0.494	0.444	172
32	DINAS PERINDUSTRIAN DAN PERDAGANGAN	0.267	0.250	0.258	32
33	DINAS PERTANIAN DAN PERKEBUNAN	0.308	0.571	0.400	28
34	DINAS PERUMAHAN RAKYAT DAN KAWASAN PERMUKIMAN	0.296	0.356	0.323	45
35	DINAS PETERNAKAN DAN KESEHATAN HEWAN	0.429	0.333	0.375	9
36	DINAS SOSIAL	0.214	0.344	0.263	122
37	DINAS TENAGA KERJA DAN TRANSMIGRASI	0.625	0.782	0.695	304
38	INSPEKTORAT	0.200	0.143	0.167	7
39	Kanwil BPN Provinsi Jawa Tengah	0.498	0.644	0.561	146
40	Kanwil Kemenag Provinsi Jawa Tengah	0.560	0.625	0.591	112
41	Kepolisian Daerah Jawa Tengah	0.567	0.576	0.571	287
42	KPU Provinsi Jawa Tengah	0.000	0.000	0.000	3
43	PLN Distribusi Jateng DIY	0.310	0.481	0.377	27

Table 12 (continued)

No.	Agency	Prec.	Rec.	F1	Support
44	Pemda Kab. Banjarnegara	0.555	0.541	0.548	98
45	Pemda Kab. Banyumas	0.456	0.587	0.513	230
46	Pemda Kab. Batang	0.574	0.586	0.580	111
47	Pemda Kab. Blora	0.663	0.712	0.687	132
48	Pemda Kab. Boyolali	0.630	0.523	0.572	130
49	Pemda Kab. Brebes	0.460	0.730	0.564	287
50	Pemda Kab. Cilacap	0.380	0.768	0.509	375
51	Pemda Kab. Demak	0.524	0.545	0.535	264
52	Pemda Kab. Grobogan	0.481	0.618	0.541	220
53	Pemda Kab. Jepara	0.617	0.580	0.598	138
54	Pemda Kab. Karanganyar	0.519	0.449	0.482	98
55	Pemda Kab. Kebumen	0.631	0.576	0.602	139
56	Pemda Kab. Kendal	0.644	0.612	0.627	170
57	Pemda Kab. Klaten	0.614	0.469	0.531	160
58	Pemda Kab. Kudus	0.618	0.596	0.607	94
59	Pemda Kab. Magelang	0.582	0.613	0.597	199
60	Pemda Kab. Pati	0.596	0.599	0.597	197
61	Pemda Kab. Pekalongan	0.713	0.618	0.662	144
62	Pemda Kab. Pemalang	0.670	0.703	0.686	192
63	Pemda Kab. Purbalingga	0.660	0.683	0.671	101
64	Pemda Kab. Purworejo	0.636	0.657	0.646	134
65	Pemda Kab. Rembang	0.627	0.515	0.566	66
66	Pemda Kab. Semarang	0.578	0.526	0.551	171
67	Pemda Kab. Sragen	0.642	0.532	0.582	77
68	Pemda Kab. Sukoharjo	0.677	0.617	0.646	141
69	Pemda Kab. Tegal	0.624	0.574	0.598	190
70	Pemda Kab. Temanggung	0.596	0.575	0.585	87
71	Pemda Kab. Wonogiri	0.677	0.698	0.688	63
72	Pemda Kab. Wonosobo	0.733	0.731	0.732	78
73	Pemda Kota Magelang	0.320	0.320	0.320	25
74	Pemda Kota Pekalongan	0.741	0.590	0.657	105
75	Pemda Kota Salatiga	0.565	0.500	0.531	26
76	Pemda Kota Semarang	0.453	0.717	0.555	499
77	Pemda Kota Surakarta	0.577	0.326	0.417	46
78	Pemda Kota Tegal	0.544	0.675	0.602	80
79	Pertamina UMPS IV Jateng-DIY	0.333	0.154	0.211	13
80	Perum Perhutani Divisi Regional Jawa Tengah	0.000	0.000	0.000	4
81	RSUD MARGONO SOEKARJO	0.000	0.000	0.000	4
82	RSUD MOEWARDI SURAKARTA	0.000	0.000	0.000	7
83	RSUD TUGUREJO SEMARANG	0.000	0.000	0.000	3
84	RSUP DR. KARIADI	0.000	0.000	0.000	3
85	SATPOL PP	0.000	0.000	0.000	5
86	SEKRETARIAT DPRD	0.000	0.000	0.000	1

REFERENCES

1. Yunqing Jiang, Patrick Cheong-Iao Pang, Dennis Wong, and Ho Yin Kan. Natural Language Processing Adoption in Governments and Future Research Directions: A Systematic Review. *Applied Sciences*, 13(22):12346, November 2023.
2. Ahmad Fathan Hidayatullah, Rosyzie Anna Apong, Daphne T.C. Lai, and Atika Qazi. Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets. *PeerJ Computer Science*, 9:e1312, June 2023.
3. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, February 2020.
4. Azzula Cerliana Zahro, Farrikh Alzami, Ramadhan Rakhmat Sani, Amiq Fahmi, Rama Aria Megantara, Muhammad Naufal, Harun Al Azies, and Iswahyudi Iswahyudi. Fairer Public Complaint Classification on LaporGub: Integrating XLM-RoBERTa with Focal Loss for Imbalance Data. *sinkron*, 9(4):1850–1862, October 2025.
5. Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.
6. Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China, 2019. Association for Computational Linguistics.
7. Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, October 2018.
8. Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
9. Salimkan Fatma Taskiran, Bahaeddin Turkoglu, Ersin Kaya, and Tunc Asuroglu. A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports*, 15(1):21631, July 2025.
10. Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice Loss for Data-imbalanced NLP Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online, 2020. Association for Computational Linguistics.
11. K. Ruwani M. Fernando and Chris P. Tsokos. Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, July 2022.
12. Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.
13. Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric Loss For Multi-Label Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, Montreal, QC, Canada, October 2021. IEEE.
14. Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep Long-Tailed Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, September 2023.
15. Qiong Hu, Masrah Azrifah Azmi Murad, Azreen Bin Azman, and Nurul Amelina Nasharuddin. A weighted difference loss approach for enhancing multi-label classification. *Scientific Reports*, 15(1):25052, July 2025.
16. Rojan-zaki Abdulkareem and Adnan Mohsin Abdulazeez. A Comparative Study of Multi-Class Classification based on Imbalanced Data. *Statistics, Optimization & Information Computing*, 15(5):4337–4357, November 2025.
17. Sri Winarno, Farrikh Alzami, Dewi Agustini Santoso, Muhammad Naufal, Harun Al Azies, Rivaldo Mersis Brilianto, and Kalaiarasi A/P Sonai Muthu. Efficient GRU-based Facial Expression Recognition with Adaptive Loss Selection. *Statistics, Optimization & Information Computing*, 14(6):3468–3499, November 2025.
18. Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. Label-Specific Feature Augmentation for Long-Tailed Multi-Label Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10602–10610, June 2023.
19. Baber Khalid, Shuyang Dai, Tara Taghavi, and Sungjin Lee. Label Supervised Contrastive Learning for Imbalanced Text Classification in Euclidean and Hyperbolic Embedding Spaces. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 58–67, San Ġiljan, Malta, 2024. Association for Computational Linguistics.
20. Grigorii Khvatskii, Nuno Moniz, Khoa D. Doan, and Nitesh V. Chawla. Class-aware contrastive optimization for imbalanced text classification. *Discover Data*, 3(1):27, July 2025.
21. Victor Kwaku Agbesi, Wenyu Chen, Md Altab Hossin, and Chiagoziem C. Ukwuoma. Dual-head alternating attention-based adaptive convolutional framework for improving low-resource and imbalanced text classification. *Neural Computing and Applications*, 37(33):27759–27779, November 2025.
22. Laouni Mahmoudi and Mohammed Salem. BalBERT: A New Approach to Improving Dataset Balancing for Text Classification. *Revue d'Intelligence Artificielle*, 37(2):425–431, April 2023.
23. Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy 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*, pages 757–770, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.
24. Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 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*, pages 843–857, Suzhou, China, 2020. Association for Computational Linguistics.
25. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online, 2020. Association for Computational Linguistics.
26. Thamer Almansour, Marwah M. Almasri, and Shimaa A. Nagro. Contextual Detection of Cross-Site Scripting Attacks Using RoBERTa-Based Deep Learning Model. *Statistics, Optimization & Information Computing*, 16(2), May 2026.
27. Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.
28. Ahmad Fathan Hidayatullah, Rosyzie Anna Apong, Daphne Teck Ching Lai, and Atika Qazi. Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer. *Social Network Analysis and Mining*, 15(1):30, March 2025.
29. Ahmad Fathan Hidayatullah. Code-Mixed Sentiment Analysis on Indonesian-Javanese-English Text Using Transformer Models. In *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 340–345, Yogyakarta, Indonesia, August 2024. IEEE.
30. Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Pascale Fung, and Ayu Purwarianti. IndoRobusta: Towards Robustness Against Diverse Code-Mixed Indonesian Local Languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 25–34, Online, 2022. Association for Computational Linguistics.
31. Yakobus Wiciaputra, Julio Young, and Andre Rusli. Bilingual Text Classification in English and Indonesian via Transfer Learning using XLM-RoBERTa. *International Journal of Advances in Soft Computing and its Applications*, 13(3):73–87, December 2021.
32. Leno Dwi Cahya, Ardytha Luthfiarta, Julius Immanuel Theo Krisna, Sri Winarno, and Adhitya Nugraha. Improving Multi-label Classification Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation Using IndoBERT Model. *Jurnal Nasional Teknologi dan Sistem Informasi*, 9(3):290–298, January 2024.
33. Ghinaa Zain Nabiilah and Derwin Suhartono. Personality Classification Based on Textual Data using Indonesian Pre-Trained Language Model and Ensemble Majority Voting. *Revue d'Intelligence Artificielle*, 37(1):73–81, February 2023.
34. Leonnardo Benjamin Hutama and Derwin Suhartono. Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic. *Informatica*, 46(8), November 2022.
35. Yulian Findawati, Agus Budi Raharjo, Dini Adni Navastara, Vincent Yonathan, Anak Agung Yatestha, and Diana Purwitasari. Multi-label Aspect Dangerous Speech Classification Using Keyword-Driven Ensemble Classifier on Imbalanced Data. *JOIV : International Journal on Informatics Visualization*, 9(4):1592, July 2025.
36. Agus Riyadi, Mate Kovacs, Uwe Serdült, and Victor Kryssanov. IndoGovBERT: A Domain-Specific Language Model for Processing Indonesian Government SDG Documents. *Big Data and Cognitive Computing*, 8(11):153, November 2024.
37. Narang Kim and Soongoo Hong. Automatic classification of citizen requests for transportation using deep learning: Case study from Boston city. *Information Processing & Management*, 58(1):102410, January 2021.
38. Xiaobo Tang, Hao Mou, Jiangnan Liu, and Xin Du. Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Scientific Reports*, 11(1):11849, June 2021.
39. Mahdi Hashemi. Automatic Type Detection of 311 Service Requests Based on Customer Provided Descriptions. *Applied Artificial Intelligence*, 36(1):2073717, December 2022.
40. Sheila Maulida Intani, Bahrul Ilmi Nasution, Muhammad Erza Aminanto, Yudhistira Nugraha, Nurhayati Muchtar, and Juan Intan Kanggrawan. Automating Public Complaint Classification Through JakLapor Channel: A Case Study of Jakarta, Indonesia. In *2022 IEEE International Smart Cities Conference (ISC2)*, pages 1–6, Pafos, Cyprus, September 2022. IEEE.
41. Evaristus Didik Madyatmadja, Bernardo Nugroho Yahya, and Cristofer Wijaya. Contextual Text Analytics Framework for Citizen Report Classification: A Case Study Using the Indonesian Language. *IEEE Access*, 10:31432–31444, 2022.
42. Yuanhang Wang, Yonghua Zhou, and Yiduo Mei. A joint attention enhancement network for text classification applied to citizen complaint reporting. *Applied Intelligence*, 53(16):19255–19265, August 2023.
43. Yunpeng Xiong, Guolian Chen, and Junkuo Cao. Research on Public Service Request Text Classification Based on BERT-BiLSTM-CNN Feature Fusion. *Applied Sciences*, 14(14):6282, July 2024.
44. Marco Esperança, Diogo Freitas, Pedro V. Paixão, Tomás A. Marcos, Rafael A. Martins, and João C. Ferreira. Proactive Complaint Management in Public Sector Informatics Using AI: A Semantic Pattern Recognition Framework. *Applied Sciences*, 15(12):6673, June 2025.
45. Siqi Chen, Yanling Zhang, Bin Song, Xiaojiang Du, and Mohsen Guizani. An Intelligent Government Complaint Prediction Approach. *Big Data Research*, 30:100336, November 2022.
46. Eric W.T. Ngai, Ariel K.H. Lui, and Brian C.W. Kei. Natural language processing in government applications: a literature review and a case analysis. *Industrial Management & Data Systems*, 125(6):2067–2104, May 2025.