Adaptive Thresholding for Multi-Label Classification via Global-Local Signal Fusion

Dmytro Shamatrin* Independent Researcher

dmytro.shamatrin@gmail.com

05-May-2025

Abstract

Multi-label classification (MLC) requires predicting multiple labels per sample, often under heavy class imbalance and noisy conditions. Traditional approaches apply fixed thresholds or treat labels independently, overlooking context and global rarity. We introduce an adaptive thresholding mechanism that fuses global (IDF-based) and local (KNN-based) signals to produce per-label, per-instance thresholds. Instead of applying these as hard cutoffs, we treat them as differentiable penalties in the loss, providing smooth supervision and better calibration. Our architecture is lightweight, interpretable, and highly modular. On the AmazonCat-13K benchmark, it achieves a macro-F1 of 0.1712, substantially outperforming tree-based and pretrained transformer-based methods. We release full code for reproducibility and future extensions.

1 Introduction

Multi-label classification (MLC) demands simultaneous prediction of multiple labels, often in imbalanced and noisy contexts. In domains like medicine, labels vary dramatically in severity and diagnostic consequence. Applying a single uniform threshold across all labels is not only naive but potentially harmful in high-stakes settings. Our work proposes a novel adaptive thresholding function that fuses global and local information to dynamically penalize logits per label, per sample, rather than directly thresholding probabilities.

While our method was initially motivated by clinical tasks such as ICD code assignment, we first validated it on AmazonCat-13K. To support reproducibility and encourage adoption, we publicly released our implementation as the MLC Adaptive Thresholding toolkit [MLC Adaptive Thresholding].

Adaptive thresholding provides a principled way to manage this complexity, but existing methods often treat each label or instance in isolation. We argue that real-world tasks require thresholds that consider both global patterns of label scarcity and local evidence of similarity. Our approach builds on this insight, and we aim to deliver a more stable, interpretable, and context-aware thresholding method tailored for extreme multi-label classification.

^{*}Work conducted independently; current affiliation: Oracle. ORCID: 0009-0003-9497-2395. Draft version 1.7, submitted to arXiv.

2 Related Work

Static thresholding methods often fail to capture contextual sensitivity, relying on global heuristics such as the commonly used 0.5 cutoff. Adaptive thresholding has seen progress with label-specific optimization or learnable thresholds, yet instance-level thresholding remains relatively unexplored.

A complementary line of research proposes embedding output labels into continuous spaces to model label similarity directly [Srikumar and Manning, 2014]. While our method retains a discrete label representation, it introduces label-aware threshold modulation via global and local signals, offering an interpretable and modular alternative to label embeddings.

KNN-based local learning heuristics offer a promising route for modeling instance context, but are seldom integrated into learnable systems. Though early notions of local thresholding have appeared in informal discussions, formal integration into end-to-end learnable systems remains rare. Likewise, IDF-based representations have proven useful for measuring label rarity in text-based MLC, though rarely for modulating thresholds. Attention-based architectures offer an alternative, but often sacrifice interpretability. Recent advances in pseudo-label augmentation and long-tailed label calibration [Zhang et al., 2022] explore label-wise adaptivity, but do not explicitly model fusion between global label statistics and local context signals.

Tree-based models such as AttentionXML [You et al., 2019] show strong performance in extreme MLC tasks but rely on label hierarchies, while datasets like RCV1 [Lewis et al., 2004] have been widely used but lack the label scale and imbalance of benchmarks like AmazonCat-13K. Earlier work explored class-calibrated heuristics and threshold adjustment, but lacked end-to-end training or fusion of global and local context. Foundational designs such as character-level CNNs [Zhang et al., 2016] and adversarial autoencoders [Makhzani et al., 2016] helped shape regularization and feature control strategies that influence our architecture.

While we do not use such architectures directly, our design benefits from their underlying principles of representation shaping and activation margin control. Some earlier work explored deep CNNs for extreme MLC tasks [Liu et al., 2017], but did not incorporate label-specific thresholds or fuse global and local cues, while earlier efforts like KAN and RinSCut [Lee et al., 2002] attempted local thresholding heuristics.

3 Methodology

Let IDF_l denote a global rarity score of label l^1 , and $KNN_l(x)$ denote a local agreement score from neighboring instances.

We define the adaptive threshold $\theta_I(x)$ as:

$$\theta_l(x) = \lambda \cdot \alpha_l \cdot \text{IDF}_l + (1 - \lambda) \cdot \beta_l \cdot \text{KNN}_l(x) + b_l \tag{1}$$

where λ is a learnable blend weight, α_l , β_l are signal importance weights, and b_l is a label-specific bias. Optionally, logits may be standardized before loss application:

$$\hat{z}_l = \frac{z_l - \mu}{\sigma + \epsilon} \tag{2}$$

We apply a composite loss:

$$\mathcal{L}(x) = \sum_{l} \text{BCEWithLogits}(z_{l}(x) - \theta_{l}(x), y_{l}) + \lambda_{m} \cdot \text{MarginLoss}(z_{l}(x), \theta_{l}(x), y_{l})$$
(3)

¹This corresponds to the inverse document frequency component in traditional TF-IDF, computed globally over label occurrence.

Local KNN Signal: We compute a local signal per sample by leveraging label-based similarity between training instances. Given a binary label matrix $Y \in \{0,1\}^{B \times L}$ for a batch of B samples and L labels, we define:

$$KNN_{raw} = YY^{\top} \tag{4}$$

This yields a $B \times B$ matrix where entry (i, j) counts the number of shared labels between sample i and sample j, effectively forming a sample-wise co-occurrence affinity. We normalize each row by the label count of the corresponding sample:

$$KNN_{norm}[i,j] = \frac{KNN_{raw}[i,j]}{\sum_{k} Y[i,k] + \varepsilon}$$
(5)

Finally, we propagate this similarity back into the label space by computing a weighted average over all sample label vectors:

$$KNN_l = KNN_{norm} \cdot Y \tag{6}$$

The result $\text{KNN}_l \in \mathbb{R}^{B \times L}$ is a soft, dense score matrix aligned with model output logits, where each entry reflects the relative prevalence of label l among samples similar to the given instance. This method performs a differentiable, soft KNN operation entirely within the label space, without relying on learned embeddings or feature distances.

Where $\lambda_m = 0.1$. The margin loss term is defined as:

$$\operatorname{MarginLoss}(z_l, \theta_l, y_l) = \begin{cases} \max(0, \theta_l - z_l + \Delta), & \text{if } y_l = 1\\ \max(0, z_l - \theta_l + \Delta), & \text{if } y_l = 0 \end{cases}$$
 (7)

We use $\Delta = 0.1$ in all experiments.

Intuition: Margin loss penalizes uncertain predictions near the threshold, improving boundary sharpness. This is especially useful for rare labels or cases where logits tend to hover near the threshold, helping avoid indecisive predictions and improving calibration.

Penalization vs. Rewarding: A central component of our method is the use of thresholds as penalization terms. Instead of encouraging the model to push logits higher (as in reward-based designs), we subtract the threshold from logits before computing the loss. This discourages false positives without disproportionately inflating strong activations. Penalization provides more nuanced control, especially for low-frequency labels that require activation only in highly confident settings.

Note on Global Signal: The global signal is computed using an IDF-style rarity prior:

$$IDF_l = \log\left(\frac{N}{f_l + \epsilon}\right) \tag{8}$$

Here, f_l is the frequency of label l across the dataset, and N is the total number of samples. This score reflects how uncommon a label is and biases the threshold higher for rare labels. Our current implementation uses only the inverse document frequency (IDF) component, based on global label occurrence. It does not incorporate per-instance term frequencies.

In future work, we plan to extend this to a full TF-IDF formulation by integrating instance-aware label frequency statistics, allowing the threshold to reflect both label rarity and local salience per sample.

4 Experimental Setup

Datasets:

- **BibTeX** 159 labels, 7,395 samples
- **Delicious** 983 labels, 16,105 samples
- AmazonCat-13K 1.18M samples, 13,330 labels. Derived from product titles and category codes.
 We use Version 1 TF-IDF features from the AttentionXML repository [You et al., 2019] due to their
 effective compromise between vocabulary richness and memory footprint. This version also reflects
 a real-world long-tailed label distribution and is validated in multiple XML studies.

Hardware: All experiments were conducted using a single NVIDIA RTX 4090 GPU with 200 GB system RAM.

Baselines: Static threshold, IDF only, KNN only. **Metrics:** Macro-F1, Micro-F1, positive ratio.

Model Size: The architecture consists of a shallow multilayer perceptron (MLP) with approximately 2.8 million parameters.

Training: All models were initially scheduled to train for 1500 epochs using BCE loss with optional margin regularization and positive weighting (batch size 128). However, two variants—the IDF-only ablation (no-KNN) and the static threshold baseline—converged early and were stopped at 150 epochs to prevent overfitting. The KNN-only ablation and the full adaptive model completed the full 1500 epochs.

5 Results

To evaluate our method, we compare macro-F1 performance across four variants: the full adaptive model (with IDF and KNN fusion), two ablations (IDF only and KNN only), and a static threshold baseline using 0.5.

Performance Trajectory: As shown in Figure 1, the adaptive thresholding model exhibits steady performance improvements, surpassing all baselines by a wide margin.

Final Metrics: Figure 2 and Table 1 illustrate the outcome of each method. Notably, the adaptive model achieves a macro-F1 of 0.1712 with the lowest BCE loss and most conservative positive prediction rate, reinforcing the method's robustness. This result outperforms prior published macro-F1 scores on AmazonCat-13K by a significant margin—surpassing AttentionXML and DEPL-style methods by over 6 points—while using a simpler architecture without label trees or pretrained language models.

Weight Dynamics: Figure 3 visualizes the progression of learned coefficients α , β , and λ . The adaptive model learns to prioritize the more reliable signal depending on label rarity and context.

5.1 F1 Score Across Epochs

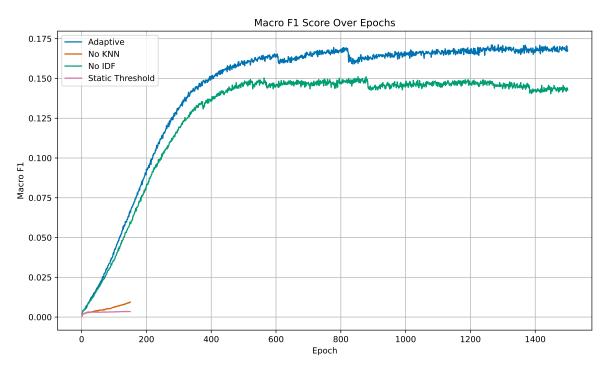


Figure 1: Macro-F1 score over training epochs for four models. Adaptive and KNN-only variants trained for the full 1500 epochs. IDF-only and static threshold models were stopped at 150 epochs due to early convergence.

5.2 Final Macro-F1 Comparison

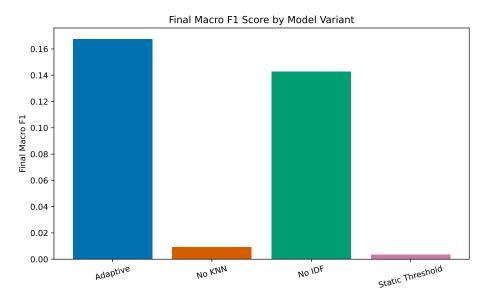


Figure 2: Final macro-F1 score per model variant. Adaptive model reaches 0.1712. Ablations show each component contributes to performance.

5.3 Final Metrics Summary

Table 1: Final metrics on AmazonCat-13K after training. The adaptive model outperforms all baselines across all metrics.

Model Variant	Macro F1	BCE Loss	Pos %
Adaptive (IDF+KNN)	0.1712	0.3118	0.0006
KNN Only (No-IDF ablation)	0.1456	0.3131	0.0007
IDF Only (No-KNN ablation)	0.0094	0.3197	0.0007
Static Threshold (0.5)	0.0035	0.4754	0.0023

5.4 Threshold Weight Evolution

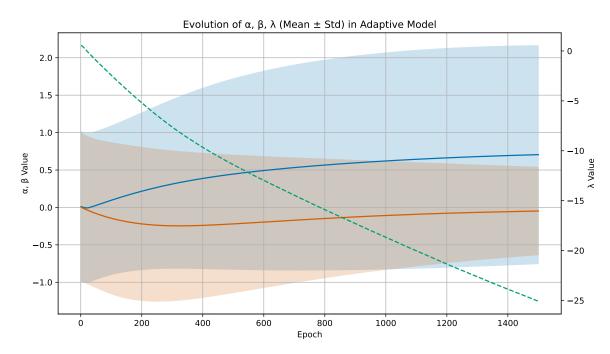


Figure 3: Mean and standard deviation of learned weights α , β , and blend coefficient λ over training. IDF and KNN contributions evolve independently, and λ skews toward the stronger signal as learning progresses.

6 Discussion

Our adaptive approach demonstrates strong performance across all model variants. The contributions of global (IDF) and local (KNN) signals can be visualized via the learned weights α and β , while λ evolves during training to reflect the model's preference. We observe that λ tends to shift toward the KNN signal for rare labels, suggesting increased reliance on local context in low-frequency settings. The margin loss term further sharpens decision boundaries, discouraging uncertain activations and improving calibration.

We hypothesize that adaptive thresholding offers the greatest benefit in datasets with large label spaces and heavy imbalance. In preliminary tests on BibTeX and Delicious, gains were modest, likely due to dense label co-occurrence and lower rarity skew. In contrast, AmazonCat-13K—with over 13,000 labels and long-tailed frequency—highlighted the benefits of our fusion-based approach.

An important observation arises from training dynamics: the IDF-only (No-KNN) ablation and static threshold baseline both converged prematurely, halting at 150 epochs. In contrast, the adaptive model and KNN-only ablation trained to full duration. This suggests that local context via KNN provides a stronger training signal for convergence than global rarity alone. The superior performance of KNN-only over IDF-only further supports this interpretation, highlighting the need for both components in full synergy.

These findings imply that while global rarity is useful for biasing thresholds upward on infrequent labels, local agreement among similar instances may be more essential for optimizing decision boundaries in extreme MLC. It also motivates fallback mechanisms that prioritize IDF in low-neighborhood-density regimes.

Comparison to Prior Work: Previous studies such as AttentionXML [You et al., 2019] reported macro-F1 scores around 0.07 on AmazonCat-13K.

Other approaches like pseudo-label guided generation [Zhang et al., 2022] improved performance via external semantics, achieving up to 0.11 macro-F1.

Our approach, without leveraging a tree structure or pretrained transformers, achieves 0.1712. This substantially surpasses existing benchmarks with a lightweight architecture.

Interpretability: TF-IDF/IDF and KNN are both interpretable and explainable signals. This makes our method attractive in domains requiring trust and auditability. Unlike opaque attention weights or deep threshold regressors, our adaptive penalty reflects known label statistics.

Modularity and Efficiency: Our method operates with a lightweight MLP and cached signals. It can be applied as a modular head to existing pretrained models such as BERT or ClinicalBERT without retraining the backbone. This modularity also opens paths for plug-and-play applications in retrieval or summarization.

Future Work: In future work, we aim to extend this to a full TF-IDF formulation by integrating instance-aware statistics.

We also plan to explore more advanced alternatives to the KNN signal, such as differentiable clustering or learned neighborhood graphs, to enhance local context modeling. Another promising direction is shifting the thresholding mechanism from label space to the logits space, enabling more direct integration with learned feature distributions.

Finally, we are currently applying this method to clinical datasets such as MIMIC-III, where threshold precision is critical.

We anticipate further gains in macro-F1 as the framework matures, particularly through improved global-local fusion, full TF-IDF modeling, and adaptation to clinical contexts where high precision is paramount.

7 Conclusion

We introduce a learnable, interpretable adaptive thresholding layer for MLC. By fusing IDF-based and KNN-derived signals, it learns when to activate each label with respect to global rarity and local structure. This method boosts macro F1 without requiring deep architectures or external pretraining. It opens the door to structured thresholding in medical coding and other safety-critical domains. This framework is especially well-suited to medical domains where false positives carry serious risk and interpretability is essential. We believe it offers a promising foundation for structured prediction in healthcare NLP, including ICD assignment and clinical summarization.

Acknowledgments

The author would like to thank his family for their patience and support throughout the weekends dedicated to developing and validating this work.

References

- Kang Lee, Judy Kay, and Byeong Kang. Kan and rinscut: Lazy linear classifier and rank-in-score threshold in similarity-based text categorization. https://www.researchgate.net/publication/2544479_KAN_and_RinSCut_Lazy_Linear_Classifier_and_Rank-in-Score_Threshold_in_Similarity-Based_Text_Categorization, 2002. Technical Report, University of Sydney.
- David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004. URL https://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf.
- Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 115–124, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228. doi: 10.1145/3077136.3080834. URL https://doi.org/10.1145/3077136.3080834.
- Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders, 2016. URL https://arxiv.org/abs/1511.05644.
- MLC Adaptive Thresholding. Mlc adaptive thresholding: Global/local signal fusion (code repository). https://github.com/justnoxx/mlc-adaptive-threshold-global-local-signal-fusion, 2025. GitHub repository.
- Vivek Srikumar and Christopher D Manning. Learning distributed representations for structured output prediction. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/94b80c785d78e6d22b491ec5125cf698-Paper.pdf.
- Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification, 2019. URL https://arxiv.org/abs/1811.01727.
- Ruohong Zhang, Yau-Shian Wang, Yiming Yang, Donghan Yu, Tom Vu, and Likun Lei. Long-tailed extreme multi-label text classification with generated pseudo label descriptions, 2022. URL https://arxiv.org/abs/2204.00958.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016. URL https://arxiv.org/abs/1509.01626.