Greedy Matching Aggregation for Soft Contrastive Learning in Multi-Label Text Classification

Ryoma Kobayashi and Jiyi Li
Hokkaido University
kobayashi.ryoma.n8@elms.hokudai.ac.jp garfieldpigljy@gmail.com

Abstract

Multi-label text classification is important for organizing large document collections, but correlated labels and skewed label distributions make representation learning difficult. Existing multi-label contrastive methods usually define pairwise supervision from label overlap, resulting in a coarse signal limited to observed shared labels. We propose Greedy Matching Aggregation Soft Contrastive Learning (GMA-SCL), which constructs contrastive targets for MLTC by greedily matching label texts and incorporating label importance weighting. The resulting reference distribution is sharpened to emphasize highly similar candidate documents while preserving graded supervision from moderately similar ones. Experiments on AAPD, Reuters-21578, and UK-LEX show that GMA-SCL improves Macro-F1 over existing multi-label contrastive baselines while maintaining competitive Micro-F1 and Hamming Loss. Our code is available at https://github.com/koba-84/gma-scl.

1 Introduction

Multi-label text classification (MLTC) assigns each document a subset of labels from a predefined label set, in contrast to multi-class classification where each instance receives exactly one label (Yang et al., 2018; Audibert et al., 2024). MLTC is widely applied to document categorization (Lewis et al., 2004), topic tagging (Yang et al., 2018), and sentiment analysis. Labels in MLTC are often correlated and the label distribution is typically long-tailed, with a few frequent labels and many rare ones (Audibert et al., 2024; Zhang et al., 2021). The combination of label correlations and long-tailed label distributions poses challenges for representation learning in MLTC.

Modern MLTC systems predominantly use pretrained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) to encode documents, followed by a linear classification head trained with binary cross-entropy (BCE) loss over sigmoid-activated logits (Yang et al., 2018). However, BCE treats labels independently and can be limited under label imbalance and label correlations (Zhang et al., 2021).

Prior work has addressed these challenges from several directions. Several approaches modify the classification objective: Asymmetric Loss (ASL) down-weights easy negative examples to handle positive-negative imbalance (Ridnik et al., 2021), while ZLPR formulates multi-label classification as a pairwise ranking problem (Su et al., 2022). Other approaches model label dependencies more explicitly, for example through autoregressive label generation (Yang et al., 2018), label correlation feedback objectives (Zhang et al., 2021), or graph-based label interaction modeling (Ma et al., 2021). These methods improve the classifier-side objective or label dependency modeling. Meanwhile, recent work has explored representation learning for MLTC through supervised contrastive learning, which uses label information to define positives and learns representations by pulling similar instances closer together while pushing dissimilar ones apart (Khosla et al., 2020; Lin et al., 2023).

In MLTC, however, positive and negative construction is ambiguous because each document can have multiple labels. Existing MLTC contrastive methods usually resolve this ambiguity through observed binary label overlap, including Jaccard-based pair weighting (Lin et al., 2023; Audibert et al., 2024) and label-specific positive sets (Zhang and Wu, 2024), with additional mechanisms for long-tailed settings (Audibert et al., 2025).

These methods motivate label-aware contrastive supervision, but their pairwise similarity signal is still derived from exact overlap between sparse label sets. This yields a coarse target signal that changes only with exactly shared labels and ignores graded label relations.

Figure 1 illustrates the intuition behind our ap-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1222–1235
November 6-10, 2026 ©2026 Association for Computational Linguistics

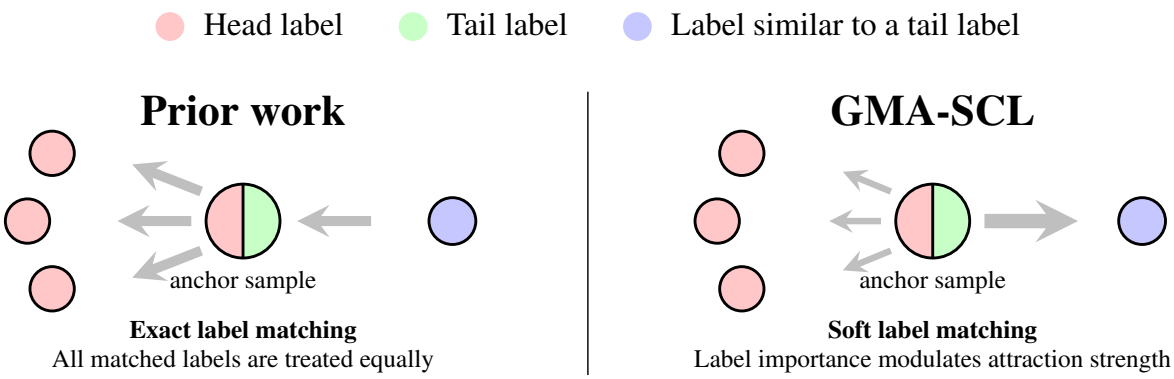


Figure 1: Contrastive supervision from binary label overlap and graded label set similarity. Each outer circle denotes a sample carrying the indicated label; the central split circle denotes the anchor sample carrying both a head and a tail label. Existing multi-label contrastive losses derive attraction from exact shared labels, which are abundant for head labels but sparse for tail labels, whereas GMA-SCL constructs a softer target from label set similarity, allowing semantically related labels, illustrated here for the tail label, to contribute to the attraction signal, with label importance modulating attraction strength.

proach. Existing multi-label contrastive objectives rely on exact label matching, so attraction is induced only by observed shared labels and matched labels are treated uniformly.

We propose Greedy Matching Aggregation Soft Contrastive Learning (GMA-SCL), which replaces exact label overlap with greedy matching based on similarities between label texts. GMA-SCL constructs a contrastive target from contextual label embeddings, label importance weighting, and greedy matching aggregation, allowing semantically related labels and more informative labels to contribute to the attraction signal.

In summary, our contributions are:

- We propose GMA-SCL, a soft contrastive target construction method for MLTC that replaces supervision based on exact label overlap with graded pairwise supervision.
- We compute the target weights using label text embeddings, label importance weighting, and greedy matching aggregation over assigned label sets.
- We evaluate GMA-SCL against end-to-end classification baselines and multi-label contrastive baselines, showing Macro-F1 improvements over contrastive baselines under frozen linear probing.
- We analyze GMA-SCL through label frequency analysis, component ablations, batch-size sensitivity, and reduced-data experiments.

2 Related Work

2.1 Contrastive Learning

Contrastive learning learns representations by pulling positive pairs together and pushing negative pairs apart. In self-supervised contrastive learning, positive pairs are usually constructed from the data itself. SimCLR treats two augmented views of the same instance as positives and contrasts them against other instances in the minibatch (Chen et al., 2020). MoCo improves contrastive learning by maintaining a dynamic dictionary with a queue and a momentum encoder (He et al., 2020).

Supervised contrastive learning uses class labels to define positives: SupCon pulls together samples from the same class and pushes apart samples from different classes (Khosla et al., 2020).

Subsequent contrastive learning methods modify different parts of this framework. Balanced contrastive learning addresses long-tailed class distributions by balancing the contribution of classes in supervised contrastive learning (Zhu et al., 2022). Prototypical contrastive learning introduces prototypes for representation learning (Li et al., 2020). Decoupled contrastive learning removes the negative-positive-coupling effect from the InfoNCE denominator (Yeh et al., 2022). CACR reweights contrastive attraction and repulsion terms (Zheng et al., 2023), while $\mathbb{X}$-CLR views contrastive learning through sample similarity graphs and replaces binary positive/negative assignments with continuous sample similarities (Sobal et al., 2025). We follow this graph perspective, but focus on constructing the target graph from multi-label

sets rather than from exact label overlap.

2.2 Multi-Label Contrastive Learning

Adapting SupCon to multi-label classification requires redefining positives, because two documents may share some positive entries in their binary label vectors but not others (Lin et al., 2023; Audibert et al., 2024). An early approach adapts supervised contrastive learning to multi-label classification by using Jaccard label-set overlap as a pairwise weighting signal (Zaigrajew and Zieba, 2022). Existing MLTC contrastive objectives address this ambiguity in several ways: weighting sample pairs by Jaccard overlap (Lin et al., 2023), constructing a separate positive set for each anchor label (Zhang and Wu, 2024), recovering label distributions from logical labels (Chen et al., 2026), or incorporating label-aware attention and label correlation into contrastive learning (Zhu et al., 2025). These studies show that label overlap and label relations provide useful contrastive supervision, but their sample-pair signals are still constructed from observed label assignments rather than from text-based label set matching.

Rare labels make this limitation more visible, because they provide only a small number of shared-label positives. MSC addresses this issue with memory, prototypes, and repulsion reweighting (Audibert et al., 2024), and a broader study reports similar limitations when positives are defined from observed shared labels (Audibert et al., 2025). This motivates a complementary direction: replacing the binary overlap signal itself with continuous label set similarity computed using label text embeddings and label importance weighting.

2.3 Text Similarity Metrics for Set-Based Soft Supervision

Automatic evaluation metrics for text generation provide several ways to compare discrete textual units. Lexical-overlap metrics such as BLEU and ROUGE measure n-gram or subsequence overlap between system outputs and references (Papineni et al., 2002; Lin, 2004). Embedding-based metrics replace exact overlap with semantic matching: BERTScore computes contextualized token similarities and aggregates them with precision, recall, and F1 variants of greedy matching (Zhang et al., 2020), while MoverScore combines contextualized embeddings with Earth Mover's Distance for softer many-to-many alignment (Zhao et al., 2019). Sentence-BERTScore shifts the comparison granularity to sentence-level semantic matching for factual consistency evaluation in summarization (Ye et al., 2024). These studies provide background for using text-based matching functions over label descriptions rather than relying only on exact label identity.

3 Method

3.1 Overview

In multi-label contrastive learning, the central design choice is how to define the supervision signal between documents. Existing objectives based on binary label overlap induce sparse and coarse graphs because they treat all shared labels equally regardless of their importance. We therefore focus on the relationships between two sets of labels and use them to construct contrastive targets for document pairs.

Specifically, GMA-SCL constructs a soft reference distribution over minibatch candidates from label set similarity. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training set, where x_i is a document and $y_i \in \{0,1\}^{|\mathcal{C}|}$ is its multi-hot label vector over the label set $\mathcal{C}$. Here, $y_{i\ell} = 1$ indicates that label ℓ is assigned to x_i. We denote the positive label set by $Y_i = \{\ell \in \mathcal{C} : y_{i\ell} = 1\}$. At each contrastive update, we sample a minibatch $B \subseteq \{1, \ldots, N\}$ and define $B \setminus \{i\}$ as the candidate set for each anchor i. The document encoder is denoted by f_θ. Let h denote the projection head. We define the normalized projected representation used in the contrastive objective as $z_i = h(f_\theta(x_i))/\|h(f_\theta(x_i))\|_2$.

Figure 2 illustrates the supervision graph used by the proposed objective: exact binary overlap produces coarse and sparse weights, whereas label set similarity induces a denser and more graded target graph.

3.2 Soft Contrastive Objective

Following the sample similarity graph view of contrastive learning (Sobal et al., 2025), we define, for each anchor i, edges to the other documents in the minibatch. Each edge (i, j) encodes the desired similarity between the two documents in the embedding space.

We first construct a soft reference graph from label set similarities. Let $\mathrm{Sim}(Y_i, Y_j)$ denote the label-set similarity between documents i and j. For each anchor i, we normalize the outgoing edges

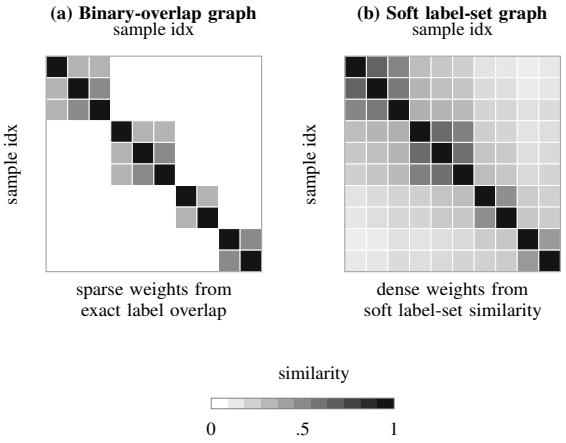


Figure 2: Sample similarity supervision graphs. Exact binary label overlap yields coarse and sparse weights, whereas GMA-SCL constructs a denser and more graded target graph from soft label set similarity.

into a reference distribution over the minibatch:

$$q_{ij} = \frac{\exp(\mathrm{Sim}(Y_i, Y_j)/\tau_q)}{\sum_{k \in B \setminus \{i\}} \exp(\mathrm{Sim}(Y_i, Y_k)/\tau_q)} \tag{1}$$

where τ_q is the label-similarity temperature.

The learned document embeddings induce a model-side distribution over the same candidate edges:

$$p_{ij} = \frac{\exp(z_i \cdot z_j/\tau_p)}{\sum_{k \in B \setminus \{i\}} \exp(z_i \cdot z_k/\tau_p)} \tag{2}$$

where τ_p is the embedding-similarity temperature.

The soft contrastive loss aligns the embedding-induced graph with this reference graph by minimizing the cross-entropy between their outgoing edge distributions:

$$\mathcal{L}^i_{\mathrm{soft}} = -\sum_{j \in B \setminus \{i\}} q_{ij} \log p_{ij} \tag{3}$$

Sharpening the similarity distributions. The label-similarity temperature τ_q controls the sharpness of the reference distribution. Following SEED (Fang et al., 2021), we use $\tau_q < \tau_p$ to sharpen the reference distribution. This concentrates probability mass on candidates in the high similarity neighborhood, making close candidates stand out clearly from less related ones while retaining nonzero supervision for moderately similar candidates.

3.3 Greedy Matching Construction of Contrastive Targets

We compute label text embeddings using a sentence embedding model g that maps each label's text to a dense vector. When a dataset provides label descriptions, $\mathrm{text}(\ell)$ is the corresponding description; otherwise, we use a simple template based on the label name, "This text is about [label name]". Concrete examples are given in Appendix B.1. As a component of label set matching, the similarity between two individual labels a and b is defined as:

$$s(a,b) = \frac{1}{2}\Big(\alpha \cdot \cos(g(\mathrm{text}(a)), g(\mathrm{text}(b))) + (1-\alpha) \cdot \mathrm{NPMI}(a,b) + 1\Big) \tag{4}$$

where $\alpha \in [0,1]$ controls the trade-off between label-text semantic similarity and label co-occurrence association. The co-occurrence term is instantiated with normalized pointwise mutual information (NPMI; Bouma, 2009), estimated from training-set label co-occurrence counts. Since both terms are bounded in $[-1,1]$, we apply an affine transformation to obtain nonnegative matching weights for the subsequent precision, recall, and harmonic-mean aggregation. Setting $\alpha = 1$ uses only label text. Details of the NPMI computation are given in Appendix C.

For label sets Y_i and Y_j, we compute their similarity by adapting the directional maximum similarity aggregation used in BERTScore (Zhang et al., 2020). With a precomputed label similarity matrix, this form compares label sets using only directional greedy matching and weighted averaging:

$$P_{ij} = \frac{\sum_{a \in Y_i} w_a \max_{b \in Y_j} s(a,b)}{\sum_{a \in Y_i} w_a}, \quad R_{ij} = \frac{\sum_{b \in Y_j} w_b \max_{a \in Y_i} s(a,b)}{\sum_{b \in Y_j} w_b} \tag{5}$$

$$\mathrm{Sim}(Y_i, Y_j) = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}} \tag{6}$$

where $w_\ell = \mathrm{idf}(\ell) = \log \frac{N}{\mathrm{freq}(\ell)}$ is the training-set IDF weight for label ℓ. This form maps each label in one set to its most similar counterpart in the other set in each direction. It produces an asymmetric pair of precision and recall scores that is combined into a harmonic-mean similarity score.

3.4 Auxiliary Ranking Regularization

The soft contrastive objective is the core training signal for GMA-SCL. We additionally include a listwise ranking term as an auxiliary regularizer. This term encourages the ordering of embedding

similarities to be consistent with the label set similarities, but it is not intended to replace the distribution matching objective. As an auxiliary objective, we add a listwise ranking loss based on ListMLE (Xia et al., 2008). Let π_i be the permutation of $B \setminus \{i\}$ that sorts $\mathrm{Sim}(Y_i, Y_j)$ in descending order, with ties broken deterministically by index. The ranking loss is defined as the negative log-likelihood of this reference ordering under the Plackett-Luce model:

$$\mathcal{L}^i_{\text{rank}} = -\sum_{r=1}^{|B|-1} \log \frac{\exp(z_i \cdot z_{\pi_i(r)}/\tau_p)}{\sum_{t=r}^{|B|-1} \exp(z_i \cdot z_{\pi_i(t)}/\tau_p)} \tag{7}$$

The joint loss is the weighted sum of the primary soft contrastive loss and the auxiliary ranking loss:

$$\mathcal{L} = \frac{1}{|B|} \sum_{i \in B} \left(\mathcal{L}^i_{\text{soft}} + \lambda \mathcal{L}^i_{\text{rank}}\right) \tag{8}$$

where $\lambda \geq 0$ is a trade-off parameter that controls the contribution of the auxiliary ranking term.

Throughout the paper, GMA-SCL denotes the proposed method family. When the auxiliary ranking term is explicitly included, we write GMA-SCL w/L_{rank}; otherwise, GMA-SCL refers to the core soft-contrastive variant.

4 Experimental Setup

This section begins with an introduction to the datasets employed in our experiments. We then describe the baselines, evaluation metrics, evaluation protocol, and implementation details.

4.1 Datasets

We evaluate on three English-language multi-label text classification benchmarks. Table 1 summarizes the dataset statistics.

- **AAPD** contains arXiv paper abstracts annotated with subject labels (Yang et al., 2018). We use the split reported by Audibert et al. (2025). We use label descriptions drawn from the publicly available arXiv Category Taxonomy[1].
- **Reuters-21578** is a newswire text categorization collection (Lewis, 1987). We use the R90 topic subset under the ModApté split (Debole and Sebastiani, 2005; Apté et al., 1994).
- **UK-LEX** contains United Kingdom legislation annotated with thematic categories from the document preamble (Chalkidis and Søgaard, 2022a,b). We use the 69-label setting and the chronological split released by Chalkidis and Søgaard (2022a).

[1] https://arxiv.org/category_taxonomy

4.2 Baselines and Metrics

We compare against classification and contrastive baselines with explicit loss definitions. Let $y_{i\ell} \in \{0, 1\}$ be the ground-truth indicator for label ℓ, $s_{i\ell}$ be the corresponding logit, and $\hat{y}_{i\ell} = \sigma(s_{i\ell})$ be the predicted label probability.

Supervised classification losses. BCE optimizes binary cross-entropy independently for each label.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|B|} \sum_{i \in B} \frac{1}{|\mathcal{C}|} \sum_{\ell \in \mathcal{C}} \Big[y_{i\ell} \log \hat{y}_{i\ell} + (1 - y_{i\ell}) \log(1 - \hat{y}_{i\ell}) \Big]. \tag{9}$$

ASL modifies BCE with asymmetric focusing terms for positive and negative labels and a hard threshold for easy negatives (Ridnik et al., 2021). Let $\hat{y}^m_{i\ell} = \max(\hat{y}_{i\ell} - m, 0)$. We set $\gamma_+ = 0$, $\gamma_- = 1$, and $m = 0$, following the ASL baseline configuration used by Audibert et al. (2025), and use the following objective:

$$\mathcal{L}_{\text{ASL}} = -\frac{1}{|B|} \sum_{i \in B} \frac{1}{|\mathcal{C}|} \sum_{\ell \in \mathcal{C}} \Big[y_{i\ell} (1 - \hat{y}_{i\ell})^{\gamma_+} \log \hat{y}_{i\ell} + (1 - y_{i\ell}) (\hat{y}^m_{i\ell})^{\gamma_-} \log(1 - \hat{y}^m_{i\ell}) \Big]. \tag{10}$$

ZLPR optimizes a zero-bounded log-sum-exp pairwise ranking loss that pushes positive logits above a fixed decision boundary and negative logits below it (Su et al., 2022). Using the positive label set Y_i and its complement $\mathcal{C} \setminus Y_i$, we define

$$\mathcal{L}_{\text{ZLPR}} = \frac{1}{|B|} \sum_{i \in B} \Big[\log \left(1 + \sum_{\ell \in Y_i} e^{-s_{i\ell}}\right) + \log \left(1 + \sum_{\ell \in \mathcal{C} \setminus Y_i} e^{s_{i\ell}}\right) \Big]. \tag{11}$$

Contrastive losses. Base is the Jaccard-based contrastive baseline and uses the normalized ob-

Table 1: Dataset statistics. # Labels denotes the number of labels. Labels/Sample and Words/Sample denote the average number of labels and words per sample, respectively. Desc. indicates whether label descriptions are available.

Dataset	Train	Val	Test	# Labels	Labels/Sample	Words/Sample	Desc.
AAPD	37,820	6,677	11,343	54	2.4	163	yes
Reuters-21578	5,827	1,943	3,019	90	1.2	128	no
UK-LEX	20,000	8,000	8,500	69	1.7	1,161	no

jective of Lin et al. (2023); Audibert et al. (2024):

$$\mathcal{L}_{\text{Base}} = -\frac{1}{|B|}\sum_{i\in B}\frac{1}{N(i)}\sum_{j\in B\setminus\{i\}}\frac{|Y_i\cap Y_j|}{|Y_i\cup Y_j|}\log p_{ij}$$
$$N(i) = \sum_{j\in B\setminus\{i\}}\frac{|Y_i\cap Y_j|}{|Y_i\cup Y_j|}. \tag{12}$$

MulSupCon instead treats each positive label of an anchor as defining a separate positive set (Zhang and Wu, 2024). For each $\ell \in Y_i$, let $P_i(\ell) = \{j \in B \setminus \{i\} : \ell \in Y_j\}$. The objective is

$$\mathcal{L}_{\text{MulSupCon}} = -\frac{1}{\sum_{i\in B}|Y_i|} \sum_{i\in B}\sum_{\ell\in Y_i}\frac{1}{|P_i(\ell)|}\sum_{j\in P_i(\ell)}\log p_{ij}. \tag{13}$$

We predict labels with a fixed sigmoid threshold of 0.5 for all methods. We report traditional metrics in MLTC, namely Micro-F1 (Mi-F1), Macro-F1 (Ma-F1), and Hamming Loss (HL) (Zhang et al., 2021).

4.3 Evaluation Protocol

We evaluate two training regimes. Classification baselines use end-to-end fine-tuning, whereas contrastive methods use contrastive pre-training followed by frozen linear probing. This protocol evaluates the representations learned by the contrastive losses, rather than the performance of fully fine-tuned downstream classifiers.

For classification baselines, we fine-tune the document encoder and the linear classification head end-to-end. For contrastive methods, we first pre-train the document encoder without text data augmentation. Following SimCLR (Chen et al., 2020), we attach a projection head h during contrastive pre-training and apply the contrastive loss to its outputs. After pre-training, we discard h and train a linear multi-label classifier on top of the frozen encoder outputs $f_\theta(x_i)$ using BCE.

Model selection and reporting are performed as follows. For the pre-training stage, the checkpoint is fixed to the final checkpoint. For each method and dataset, we select hyperparameters using validation Macro-F1 and report the corresponding test result. Main results are reported as mean and standard deviation over three random seeds.

4.4 Implementation Details

We use RoBERTa-base[2] (Liu et al., 2019) as the document encoder for AAPD and Reuters-21578, and Legal-BERT-base[3] (Chalkidis et al., 2020) for UK-LEX. We use the [CLS] token as the document representation. For contrastive methods, the projection head h is composed of two fully connected layers with ReLU as the activation function, $h(\cdot) = W_2\text{ReLU}(W_1\cdot)$, where $W_1 \in \mathbb{R}^{d_h\times 768}$ and $W_2 \in \mathbb{R}^{d\times d_h}$. We set $d_h = 512, d = 128$ in our experiments. Inputs to both the document encoder and the label encoder were tokenized with a maximum sequence length of 512 tokens, truncating longer inputs. For the label encoder, we use the frozen all-MiniLM-L6-v2[4] model from Sentence-Transformers (Reimers and Gurevych, 2019). We encode each label text once before training, precompute and cache the label similarity matrix, and use it to construct minibatch reference distributions during contrastive pre-training, so the training loop does not require an additional forward pass through the label encoder. We fix the embedding-similarity temperature to $\tau_p = 0.1$ in all experiments.

We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 0.01, a batch size of 64, and bfloat16 mixed precision (Kalamkar et al., 2019). For contrastive pre-training, we use linear warmup for the first 5% of training, followed by cosine annealing. For end-to-end classification

[2] https://huggingface.co/FacebookAI/roberta-base

[3] https://huggingface.co/nlpaueb/legal-bert-base-uncased

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

and linear probing, we use linear decay to zero without warmup. Hyperparameter candidate sets, software versions, parameter counts, and compute budget are reported in Appendix B. Contrastive pre-training is run for 80 epochs, and both end-to-end classification and linear probing are run for 40 epochs.

5 Results and Analysis

5.1 Main Results

Table 2 presents the main results on the three benchmark datasets. The classification baselines are trained by end-to-end fine-tuning without contrastive objectives, while the contrastive methods use contrastive pre-training followed by frozen linear probing.

The proposed GMA-SCL family achieves the strongest Macro-F1 results overall across the three datasets. The end-to-end classification baselines remain competitive, showing that end-to-end fine-tuning is a strong supervised setting. Compared with the contrastive baselines, the proposed GMA-SCL family shows consistent Macro-F1 improvements, suggesting that graded label-set supervision provides a useful contrastive signal under the evaluated protocol. For Micro-F1 and Hamming Loss, the proposed GMA-SCL family remains competitive with generally modest margins, while achieving the best values in several dataset-metric settings. The relatively small margin over Base on Reuters-21578 may reflect its low label cardinality, with only 1.2 labels per sample on average. The ranking term is not uniformly beneficial across datasets, potentially because the ranking and soft contrastive losses have substantially different scales and the selected coefficient may not balance them equally.

5.2 Label Frequency Analysis

Figure 3 analyzes label classification performance across different frequency groups on AAPD. Following Zhang et al. (2021), we divide all labels into four groups according to their frequency in the AAPD training set, from head labels to tail labels.

Figure 3(a) illustrates the label frequency distribution on the AAPD training set, which exhibits a heavily long-tailed distribution. Figure 3(b) illustrates the Macro-F1 performance of different methods in the four groups.

As shown in Figure 3(b), performance generally decreases as label frequency decreases, confirming that rare labels are harder to classify. At the same time, GMA-SCL shows its largest advantage in the tail group. This pattern suggests that soft label set similarity is especially useful for rare labels, where exact shared-label signals are sparse.

5.3 Ablation Study

We conduct a component ablation of the core GMA-SCL objective on AAPD. We compare the core configuration, a variant without reference sharpening, and a variant without IDF weighting, and a variant replacing the proposed label set similarity with Jaccard overlap.

Table 3 reports the ablation results. Removing reference sharpening causes the largest degradation, suggesting that it is important for the soft target distribution. Removing IDF weighting leads to a small decrease in Macro-F1. The Jaccard variant remains close to the core and improves over Base. Overall, the ablation supports the combination of reference sharpening and semantically informed label-set similarity as the central design of GMA-SCL.

5.4 Batch Size Sensitivity

Batch size is an important factor in contrastive learning, since batch-wise contrastive objectives use other samples in the minibatch as comparison samples. We therefore examine whether the advantage of GMA-SCL over the Base contrastive baseline is preserved when the contrastive pre-training batch size is varied. Following SimCLR (Chen et al., 2020), we linearly scale the contrastive pre-training learning rate with the batch size. All other settings, including linear probing, are kept fixed.

Figure 4 reports the Δ Macro-F1 of GMA-SCL over Base. The gain remains positive across all tested batch sizes, with larger observed gains at smaller batch sizes. This diagnostic analysis suggests that GMA-SCL preserves its advantage over Base even when the minibatch provides fewer comparison samples.

5.5 Data Efficiency

This analysis examines the data efficiency of GMA-SCL with varying fractions of the training data. We compare with Base, the closest in-batch contrastive baseline, to isolate the effect of replacing Jaccard-based supervision with soft label set similarity.

We reduce only the AAPD training split to 5%, 10%, 20%, and 50% using iterative multi-label stratification (Sechidis et al., 2011), with the original training split serving as the 100% setting, while

Table 2: Main results on all benchmarks (mean ± std over three seeds; HL is multiplied by 10^3). Best results are in **bold** within each dataset.

	AAPD			Reuters-21578			UK-LEX		
Method	Mi-F1 ↑	Ma-F1 ↑	HL ↓	Mi-F1 ↑	Ma-F1 ↑	HL ↓	Mi-F1 ↑	Ma-F1 ↑	HL ↓
BCE	72.35 ± 0.19	57.85 ± 0.14	23.79 ± 0.26	90.20 ± 0.22	57.20 ± 0.76	2.65 ± 0.06	72.25 ± 0.33	61.70 ± 0.57	12.58 ± 0.26
ASL	72.81 ± 0.11	58.77 ± 0.33	23.82 ± 0.04	90.54 ± 0.23	60.62 ± 1.46	2.60 ± 0.05	72.09 ± 0.66	61.92 ± 0.78	12.79 ± 0.27
ZLPR	72.54 ± 0.19	58.77 ± 0.13	23.88 ± 0.48	90.66 ± 0.11	58.44 ± 0.81	2.52 ± 0.03	72.88 ± 0.48	62.36 ± 0.31	**12.14** ± 0.34
Base	73.17 ± 0.07	59.15 ± 0.12	23.73 ± 0.07	90.68 ± 0.11	62.74 ± 1.62	2.57 ± 0.02	72.91 ± 0.02	61.96 ± 0.39	12.35 ± 0.06
MulSupCon	73.21 ± 0.13	59.46 ± 0.29	23.60 ± 0.28	90.70 ± 0.13	61.14 ± 0.85	2.55 ± 0.05	73.00 ± 0.34	62.11 ± 0.66	12.21 ± 0.09
GMA-SCL	73.47 ± 0.20	60.19 ± 0.10	23.58 ± 0.21	**90.95** ± 0.18	**64.34** ± 1.54	**2.49** ± 0.05	**73.10** ± 0.02	**63.03** ± 0.38	12.38 ± 0.06
GMA-SCL w/L_{rank}	**73.54** ± 0.11	**60.26** ± 0.33	**23.46** ± 0.09	90.74 ± 0.12	63.00 ± 0.36	2.55 ± 0.03	73.05 ± 0.27	62.68 ± 0.43	12.35 ± 0.19

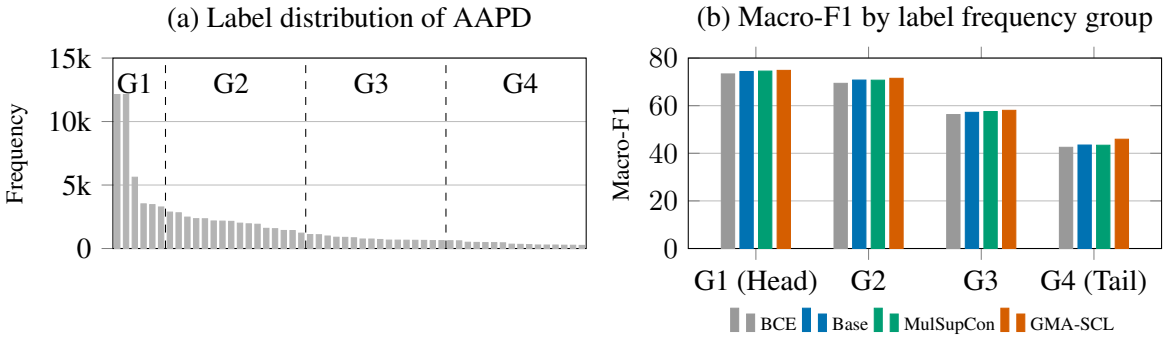


Figure 3: Label frequency analysis on AAPD. (a) Training set label distribution with four frequency groups (G1: head, G4: tail) separated by dashed lines. (b) Per-group Macro-F1 scores averaged over the same three random seeds as the main results.

Table 3: Component ablations of the core GMA-SCL objective on AAPD (averaged over three seeds; HL is multiplied by 10^3).

	Mi-F1 ↑	Ma-F1 ↑	HL ↓
GMA-SCL	73.47	60.19	23.58
w/o sharpening	73.03	59.21	23.79
w/o IDF weighting	73.44	59.93	23.53
Jaccard similarity	73.48	59.84	23.58

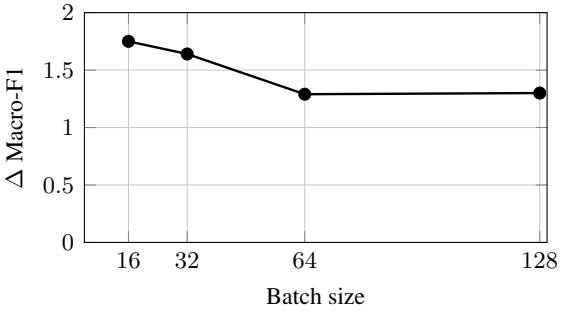


Figure 4: Batch size sensitivity on AAPD. Δ Macro-F1 is GMA-SCL minus Base. Results are from a single run per batch size.

keeping the validation and test splits fixed. For each fraction, we generate one stratified subset and reuse it across both methods. We select the linear probing learning rate from three candidates on the validation split and report the corresponding test result from a single run.

Figure 5 shows that GMA-SCL remains above Base across all training fractions, suggesting improved data efficiency over Jaccard-based supervision under this reduced-data protocol.

6 Conclusion

We proposed GMA-SCL, which constructs soft contrastive targets for MLTC using graded pairwise supervision instead of exact label overlap. GMA-SCL combines label text embeddings, label importance weighting, greedy matching aggregation, and reference sharpening to define target similarities between multi-label documents. Experiments across three benchmark datasets show that GMA-SCL improves Macro-F1 over existing multi-label contrastive baselines under frozen linear probing

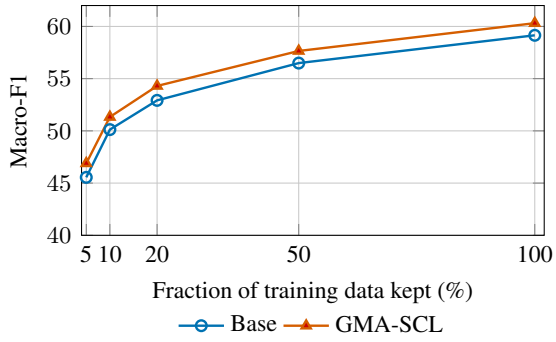


Figure 5: Data efficiency analysis on AAPD. Macro-F1 is reported after selecting the linear probing learning rate on the validation split. Results use one stratified subset and one run per fraction.

while maintaining competitive Micro-F1 and Hamming Loss. These results suggest that the proposed target construction can provide useful graded supervision under the evaluated protocol. Future work will explore using large language models to enrich underspecified label texts.

Limitations

GMA-SCL relies on textual label information to construct the semantic component of label set similarity. When label descriptions are unavailable, underspecified, or weakly aligned with document semantics, the resulting reference distribution may be less reliable.

Our experiments are limited to English-language MLTC benchmarks with moderate label spaces. Further work is needed to verify whether the same conclusions hold for multilingual datasets, extreme multi-label classification, and domains with larger or noisier label vocabularies.

Finally, the method introduces additional label set matching overhead within each minibatch. Although this cost is modest in our experiments, larger batches or larger label spaces may require more efficient caching, approximation, or matching strategies.

Ethics Statement

This work studies multi-label text classification on public research datasets. Potential positive impacts include improving the organization, retrieval, and analysis of large document collections. However, the proposed method should be viewed as decision support rather than a fully automated decision system.

Incorrect labels may hide relevant documents, surface irrelevant ones, or propagate wrong metadata to downstream systems. These risks are more serious in legal-domain applications, where misclassification may affect search, screening, or compliance analysis. The method may also inherit biases from long-tailed label distributions and from incomplete or ambiguous label texts, which can lead to underrepresentation of rare or specialized categories. We therefore recommend domain-specific validation, label-frequency analysis, and human review before using such models in consequential applications.

We used AI assistants for language polishing, editing, and preliminary literature search assistance. All cited literature, scientific claims, experiments, and analyses were verified by the authors.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K28092.

References

Chidanand Apté, Fred Damerau, and Sholom M. Weiss. 1994. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251.

Alexandre Audibert, Aurélien Gauffre, and Massih-Reza Amini. 2024. Exploring contrastive learning for long-tailed multi-label text classification. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part VII*, pages 245–261, Berlin, Heidelberg. Springer-Verlag.

Alexandre Audibert, Aurélien Gauffre, and Massih-Reza Amini. 2025. Multi-label contrastive learning : A comprehensive study. *Preprint*, arXiv:2412.00101.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference 2009*, Potsdam, Germany. GGSCL.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis and Anders Søgaard. 2022a. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454,

Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis and Anders Søgaard. 2022b. UK-LEX Dataset - Part of Chalkidis and Søgaard (2022).

Ning Chen, Shen-Huan Lyu, Tian-Shuang Wu, Yanyan Wang, and Bin Tang. 2026. Improving multi-label contrastive learning by leveraging label distribution. *Pattern Recognition*, 174:113011.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Franca Debole and Fabrizio Sebastiani. 2005. An analysis of the relative hardness of Reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, 56(6):584–596.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. 2021. SEED: Self-supervised distillation for visual representation. In *International Conference on Learning Representations*.

Patricia Fischer, Sebastian Pütz, and Daniël de Kok. 2019. Association metrics in neural transition-based dependency parsing. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 181–189, Paris, France. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. 2019. A study of BFLOAT16 for deep learning training. *Preprint*, arXiv:1905.12322.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

David Lewis. 1987. Reuters-21578 Text Categorization Collection. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52G6M.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2020. Prototypical contrastive learning of unsupervised representations. *Preprint*, arXiv:2005.04966.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3855–3864, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2011*, volume 6913 of *Lecture Notes in Computer Science*, pages 145–158. Springer.

Vlad Sobal, Mark Ibrahim, Randall Balestriero, Vivien Cabannes, Diane Bouchacourt, Pietro Astolfi, Kyunghyun Cho, and Yann LeCun. 2025. 𝕏-sample contrastive loss: Improving contrastive learning with sample similarity graphs. In *The Thirteenth International Conference on Learning Representations*.

Jianlin Su, Mingren Zhu, Ahmed Murtadha, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2022. Zlpr: A novel loss for multi-label classification. *Preprint*, arXiv:2208.02955.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *International Conference on Machine Learning*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuxuan Ye, Edwin Simpson, and Raul Santos Rodriguez. 2024. Using similarity to evaluate factual consistency in summaries. *arXiv preprint arXiv:2409.15090*.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. 2022. Decoupled contrastive learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 668–684, Berlin, Heidelberg. Springer-Verlag.

Vladimir Zaigrajew and Maciej Zieba. 2022. Contrastive learning for multi-label classification. In *Proceedings of conference on neural information processing systems, New Orleans*, pages 1–8.

Pingyue Zhang and Mengyue Wu. 2024. Multi-label supervised contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:16786–16793.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ximing Zhang, Qian-Wen Zhang, Zhao Yan, Ruifang Liu, and Yunbo Cao. 2021. Enhancing label correlation feedback in multi-label text classification via multi-task learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1190–1200, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Huangjie Zheng, Xu Chen, Jiangchao Yao, Hongxia Yang, Chunyuan Li, Ya Zhang, Hao Zhang, Ivor Tsang, Jingren Zhou, and Mingyuan Zhou. 2023. Contrastive attraction and contrastive repulsion for representation learning. *Transactions on Machine Learning Research*.

Bo Zhou, Liulei Li, Yujia Wang, Huafeng Liu, Yazhou Yao, and Wenguan Wang. 2025. Unialign: Scaling multimodal alignment within one unified model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29644–29655.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. 2022. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6908–6917.

Zhengzhong Zhu, Pei Zhou, Zeting Li, Kejiang Chen, and Jiangping Zhu. 2025. Multi-label text classification with label attention aware and correlation aware contrastive learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 8420–8428. International Joint Conferences on Artificial Intelligence Organization. Main Track.

A Broader Applicability

Since GMA-SCL compares label sets through semantic label embeddings and BERTScore-like matching rather than exact label identities, it can potentially extend to heterogeneous label spaces without manually specified cross-dataset or cross-modal label correspondences, provided that textual label descriptions are available, as explored from a related perspective in UNIALIGN (Zhou et al., 2025).

B Additional Experimental Details

Table 4 reports the candidate sets for the tuned hyperparameters. Table 5 reports the selected values used for the reported test results. For τ_q, the candidate set includes $\tau_q = \tau_p = 0.10$, which corresponds to a non-sharpened reference distribution. For α, validation selection chooses $\alpha = 1.0$ on AAPD and $\alpha = 0.5$ on Reuters-21578 and UK-LEX. Thus, the NPMI component is dataset dependent, and we treat it as an optional empirical association signal rather than a required component of GMA-SCL.

B.1 Label Text Examples

Table 6 shows examples of label texts used to compute label embeddings. AAPD uses label descriptions from the arXiv Category Taxonomy, whereas Reuters-21578 and UK-LEX use a template based on the label name because label descriptions are unavailable.

B.2 Compute environment and model size

All experiments use the Hugging Face implementations of the document encoders and the Sentence-Transformers label encoder. Each experiment was run on a single NVIDIA A100 80GB PCIe GPU with bfloat16 mixed precision. The total computational budget for all experiments and evaluations was roughly 2,000 GPU hours.

The document encoder has 125M parameters for RoBERTa-base and 110M parameters for Legal-BERT-base-uncased. The frozen label text encoder, Sentence-Transformers all-MiniLM-L6-v2, has 22.7M parameters and is used only to precompute label embeddings. These counts refer to the pretrained encoder backbones.

B.3 Software versions

Table 7 reports the software package versions used in the experiments.

B.4 Artifact licenses and intended use

Reuters-21578 and UK-LEX are released under CC BY 4.0, while the AAPD processed split used in this work does not provide an explicit license. RoBERTa-base, Legal-BERT-base-uncased, and all-MiniLM-L6-v2 are released under MIT, CC BY-SA 4.0, and Apache-2.0 licenses, respectively, according to their public model cards. We use these third-party artifacts only for research experiments and do not redistribute the datasets or pretrained checkpoints.

B.5 Data privacy and content.

We use existing public benchmark datasets and do not collect new user data or create new annotations. Reuters-21578 and UK-LEX may contain public names, incidental identifiers, or offensive content from naturally occurring newswire and legal documents, but these are not prediction targets. We do not identify individuals, infer personal attributes, release document-level predictions, or publish extracted identifiers. No additional anonymization was performed because we do not create or redistribute a new raw-text dataset.

C NPMI Component

We estimate empirical label occurrence and co-occurrence probabilities from the training split. For observed label pairs with $p(a, b) > 0$, we compute

$$\mathrm{NPMI}(a, b) = \frac{\log \frac{p(a,b)}{p(a)p(b)}}{-\log p(a, b)}. \qquad (14)$$

For identical labels, we set $\mathrm{NPMI}(a, a) = 1$.

For unobserved label pairs, we assign a neutral default value of 0, following prior work that maps entries absent from an association score table to the neutral value (Fischer et al., 2019). This is not clipping: observed pairs retain their computed NPMI values, including negative values.

Table 4: Hyperparameter candidate sets. Shared candidate sets are used for all datasets, while dataset-specific candidate sets are tuned separately for each dataset. Selection is based on validation Macro-F1.

Hyperparameter / stage	AAPD	Reuters-21578	UK-LEX
End-to-end classification LR	$\{5 \times 10^{-5}, 1 \times 10^{-4}\}$	$\{5 \times 10^{-5}, 1 \times 10^{-4}\}$	$\{5 \times 10^{-5}, 1 \times 10^{-4}\}$
Contrastive pre-training LR	$\{5 \times 10^{-5}, 1 \times 10^{-4}\}$	$\{5 \times 10^{-5}, 1 \times 10^{-4}\}$	$\{5 \times 10^{-5}, 1 \times 10^{-4}\}$
Linear probe LR	$\{5 \times 10^{-4}, 1 \times 10^{-3}\}$	$\{5 \times 10^{-3}, 1 \times 10^{-2}\}$	$\{5 \times 10^{-3}, 1 \times 10^{-2}\}$
τ_q	$\{0.01, 0.04, 0.07, 0.10\}$	$\{0.01, 0.04, 0.07, 0.10\}$	$\{0.01, 0.04, 0.07, 0.10\}$
α	$\{0.5, 1.0\}$	$\{0.5, 1.0\}$	$\{0.5, 1.0\}$
λ	$\{0, 10^{-4}, 10^{-3}\}$	$\{0, 10^{-4}, 10^{-3}\}$	$\{0, 10^{-4}, 10^{-3}\}$

Table 5: Selected hyperparameter values used for the reported test results. All values are selected on the validation split using Macro-F1.

Method	Hyperparameter	AAPD	Reuters-21578	UK-LEX
End-to-end classification				
BCE	LR	5×10^{-5}	1×10^{-4}	1×10^{-4}
ASL	LR	5×10^{-5}	1×10^{-4}	1×10^{-4}
ZLPR	LR	5×10^{-5}	5×10^{-5}	5×10^{-5}
Contrastive methods				
Base	Contrastive LR	5×10^{-5}	5×10^{-5}	5×10^{-5}
	Linear-probe LR	1×10^{-3}	1×10^{-2}	1×10^{-2}
MulSupCon	Contrastive LR	5×10^{-5}	5×10^{-5}	5×10^{-5}
	Linear-probe LR	1×10^{-3}	1×10^{-2}	5×10^{-3}
GMA-SCL	Contrastive LR	5×10^{-5}	5×10^{-5}	5×10^{-5}
	Linear-probe LR	5×10^{-4}	1×10^{-2}	5×10^{-3}
	τ_q	0.01	0.04	0.04
	α	1.0	0.5	0.5
GMA-SCL w/L_{rank}	Contrastive LR	5×10^{-5}	5×10^{-5}	5×10^{-5}
	Linear-probe LR	5×10^{-4}	1×10^{-2}	5×10^{-3}
	τ_q	0.01	0.04	0.04
	α	1.0	0.5	0.5
	λ	10^{-4}	10^{-4}	10^{-3}

Dataset	Label name	Label text
AAPD	cs.AI	Covers all areas of AI except Vision, Robotics, Machine Learning, Multiagent Systems, and Computation and Language (Natural Language Processing), which have separate subject areas. In particular, includes Expert Systems, Theorem Proving (although this may overlap with Logic in Computer Science), Knowledge Representation, Planning, and Uncertainty in AI. Roughly includes material in ACM Subject Classes I.2.0, I.2.1, I.2.3, I.2.4, I.2.8, and I.2.11.
Reuters-21578	acq	This text is about acq
UK-LEX	AGRICULTURE	This text is about AGRICULTURE

Table 6: Examples of label texts used to compute label embeddings. AAPD uses label descriptions from the arXiv Category Taxonomy, while Reuters-21578 and UK-LEX use the template because label descriptions are unavailable.

Package	Version
Python	3.12.3
PyTorch	2.6.0+cu124 (CUDA 12.4)
Lightning	2.6.1
TorchMetrics	1.9.0
Transformers	4.57.6
Tokenizers	0.22.2
Datasets	2.14.3
Sentence-Transformers	5.3.0

Table 7: Software package versions used in the experiments.