Multi-Label Text Classification with Label Attention Aware and Correlation Aware Contrastive Learning

Zhengzhong Zhu[1] , **Pei Zhou**[1] , **Zeting Li**[1] , **Kejiang Chen**[1] , **Jiangping Zhu**[1,*]
[1]College of Computer Science, Sichuan University, Chengdu, China
{zjp16,ZhengzhongZhu}@scu.edu.cn

Abstract

Multi-label text classification (MLTC) is a challenging task where each document can be associated with multiple interdependent labels. This task is complicated by two key issues: the intricate correlations among labels and the partial overlap between labels and text relevance. Existing methods often fail to capture the semantic dependencies between labels or struggle to handle the ambiguities caused by partial overlaps, resulting in suboptimal representation learning. To address these challenges, we propose the Unified Contextual and Label-Aware Framework (UCLAF), which integrates a Label Attention Aware Network(LAN) and Correlation Aware Contrastive Learning (CACL) in a synergistic design. The Label Attention Aware Network explicitly models label dependencies by embedding labels and texts into a shared semantic space, aligning text representations with label semantics. Meanwhile, Correlation Aware Contrastive Learning refines these representations by dynamically modeling sample-level relationships, leveraging a contrastive loss function that accounts for the proportional overlap of labels between samples. This complementary approach enables UCLAF to jointly address complex label correlations and partial label overlaps. Extensive experiments on benchmark datasets demonstrate that UCLAF significantly outperforms state-of-the-art methods, showcasing its effectiveness in improving both representation learning and classification performance in MLTC tasks. We will release our code after the paper is accepted.

1 Introduction

Multi-label text classification (MLTC) is a critical task in natural language processing (NLP), where each document can be associated with multiple relevant labels simultaneously. This task underpins numerous real-world applications, such as document categorization [Rubin *et al.*, 2012], news tagging [Zhang *et al.*, 2019], and research topic classification [Yang *et al.*, 2016]. Compared to single-label classification, MLTC introduces unique challenges due to the inherent complexities of label relationships and overlaps.

One of the primary challenges in MLTC lies in the complex correlations among labels. Labels in MLTC often exhibit hierarchical, semantic, or co-occurrence dependencies. For example, an article discussing climate change may be tagged with “Environment,” “Policy,” and “Economics.” These labels are semantically related, as climate change often involves policy discussions and economic impacts. Ignoring such intricate correlations risks producing incoherent predictions and undermines the interpretability of the model. Additionally, MLTC often faces the issue of partial label overlaps, where a document aligns strongly with some labels but only partially with others. For instance, an article on renewable energy may be highly relevant to “Environment” and “Energy” while weakly overlapping with “Economics” when financial aspects are discussed. These challenges are illustrated in Table 1.

Text	**Labels**
Global leaders discussed the economic impacts of climate change and the urgent need for policy reforms to address environmental challenges.	Environment, Policy, Economics
The impact of renewable energy policies on reducing carbon emissions has been widely discussed, with a focus on solar and wind energy integration.	Environment, Energy, Policy, Economics
Advancements in deep learning have significantly improved autonomous driving, enabling better perception and decision-making capabilities.	AI, Robotics, Transportation, Ethics
The study explores the ethical implications of using AI in healthcare for disease diagnosis and treatment planning.	AI, Healthcare, Ethics
This review examines smartphone features such as battery performance, camera quality, and overall cost efficiency.	Technology, Performance, Price

Table 1: An example of several papers from arXiv.

Although existing methods address parts of these challenges, limitations remain. Classical approaches, such as binary relevance and label powerset, treat labels as independent entities, completely ignoring label correlations and overlaps [Guo *et al.*, 2021; Wang *et al.*, 2022]. More recent methods, such as label embedding and latent space learning, model label correlations to some extent [Ying *et al.*, 2021; Xu *et al.*, 2023], but fail to handle the ambiguity caused by partial overlaps. These approaches often treat overlapping labels as either fully dependent or entirely unrelated, leading to suboptimal representation learning and reduced generalization capabilities.

Motivated by these challenges, we propose a unified frame-

[*] Corresponding author

work to explicitly address both complex label correlations and partial label overlaps in MLTC. Our solution is rooted in the observation that these two challenges are inherently interconnected and require complementary solutions. We designed the Unified Contextual and Label-Aware Framework (UCLAF), which integrates two key components: (1) the Label Attention Aware Network(LAN), which explicitly captures label dependencies by embedding text and labels into a shared semantic space; and (2) the Correlation Aware Contrastive Learning (CACL) , which introduces a dynamic and fine-grained mechanism for defining and optimizing relationships between samples based on their label sets. Unlike traditional methods[Zhang and Wu, 2024; Khosla *et al.*, 2020] that rely on strict label matching (e.g., exact matches or any overlap), CACL considers the proportional overlap of labels between samples to determine their similarity and effectively mitigates the ambiguities caused by partial label overlaps and enhances representation learning in multi-label contexts. Two components form a complementary and synergistic design. Label Attention Module explicitly models label correlations by aligning text and label representations in a shared semantic space, CACL refines these representations by focusing on inter-sample relationships. This synergy ensures a comprehensive solution to the challenges of multi-label text classification. In summary, the main contributions of this work are as follows:

- We identify partial label overlaps as a key scientific problem in MLTC and analyze its impact on positive-negative sample distinctions and representation learning.
- We propose UCLAF, a unified framework that integrates label-aware representation learning and supervised contrastive learning to address partial label overlaps and model complex label correlations.
- Extensive experiments on benchmark datasets, including AAPD, EURLex demonstrate that UCLAF significantly outperforms state-of-the-art methods in multiple evaluation metrics.

2 Related work

2.1 Multi-Label Learning

Multi-label learning deals with instances associated with multiple interrelated labels and finds applications in domains such as object detection and anomaly detection [Ge *et al.*, 2018; Cheng *et al.*, 2025], image classification [Zhang and Wu, 2024], and text classification [Xiao *et al.*, 2019]. For example, an image may contain several objects, or a text may cover multiple topics [Bogatinovski *et al.*, 2022; Liu *et al.*, 2017a; Liu *et al.*, 2021; Wu *et al.*, 2020]. Early methods approached this by treating multi-label problems as independent single-label tasks, which failed to model label correlations effectively [Boutell *et al.*, 2004; Read *et al.*, 2011]. Recent advances utilize statistical analysis [Wang *et al.*, 2018; Yeh *et al.*, 2017], label embedding techniques like RNNs and GNNs [Yazici *et al.*, 2020; Wang *et al.*, 2016; Durand *et al.*, 2019], or auto-encoders to jointly encode label and instance correlations [Zhao *et al.*, 2021; Bai *et al.*, 2020; Bai *et al.*, 2022]. Despite progress, most work focuses on non-text domains, leaving textual multi-label learning relatively underexplored.

2.2 Multi-label Text Classification

Existing multi-label text classification (MLTC) methods primarily focus on learning text representations and modeling label correlations. Early approaches used CNNs[Moschitti *et al.*, 2014; Kurata *et al.*, 2016] and RNNs[Liu *et al.*, 2016] to capture long-range dependencies in text. We propose that focusing on each label's unique feature representation is a promising strategy for obtaining richer text representations for each label. For example, recent methods[Wang *et al.*, 2023] have incorporated Transformer-based modules to exploit label semantic for capturing high-quality document representations, while long-range word dependencies are modeled through encoders, and multi-label attention mechanisms identify[Xu *et al.*, 2023] the most relevant parts of the text for each label. In terms of label correlation modeling, previous approaches have employed sequence generation models[Nam *et al.*, 2017; Yang *et al.*, 2018; Xiao *et al.*, 2021] and iterative inference mechanisms[Wang *et al.*, 2021]. Additionally, Ma et al.[Ma *et al.*, 2021] applied graph neural networks based on labeled graphs. However, these methods have yet to fully explore the latent semantic space between labels and texts, and the challenge of obtaining more effective feature representations remains a significant issue in MLTC.

In light of these challenges, our paper is specifically designed for multi-label text classification (MLTC). This approach aims to address the existing limitations of complex label correlations and partial overlaps through the integration of a unified framework.

3 Task Formulation

Consider a batch of data $B = \{(doc_i, y_i)\}_{i=1}^{N}$, where N represents the mini-batch size, and $y_i = \{y_j^{(i)}\}_{j\in\{0,1\}}^{L}$ denotes the multi-label set for sample i. Here, $y_j^{(i)}$ indicates the j-th label of sample i, and L is the total number of labels in the dataset. The classifier computes the probability $\hat{y}_i = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L\}$, where each $\hat{y}_j$ represents the likelihood of the corresponding label being true.

The binary cross-entropy (BCE) loss is used to measure the discrepancy between the predicted probabilities $\hat{y}_i$ and the ground truth labels y_i. It is defined as:

$$\mathcal{L}_{\text{BCE}}(\hat{y}_i, y_i) = -\frac{1}{L}\sum_{l\in L}\left[y_l \log \hat{y}_l + (1-y_l)\log(1-\hat{y}_l)\right], \tag{1}$$

where y_l and $\hat{y}_l$ correspond to the ground truth and predicted probability for label l, respectively.

For the contrastive learning framework, which is based on the MoCo structure, we introduce additional notations. Let $z_i^{(q)}$ and $z_i^{(k)}$ denote the L2-normalized outputs of the query model and key model for sample i, respectively. The key model is updated using momentum-based updates. Furthermore, a queue Q is maintained to store $z^{(k)}$ from previous batches, following the MoCo[He *et al.*, 2020] design.

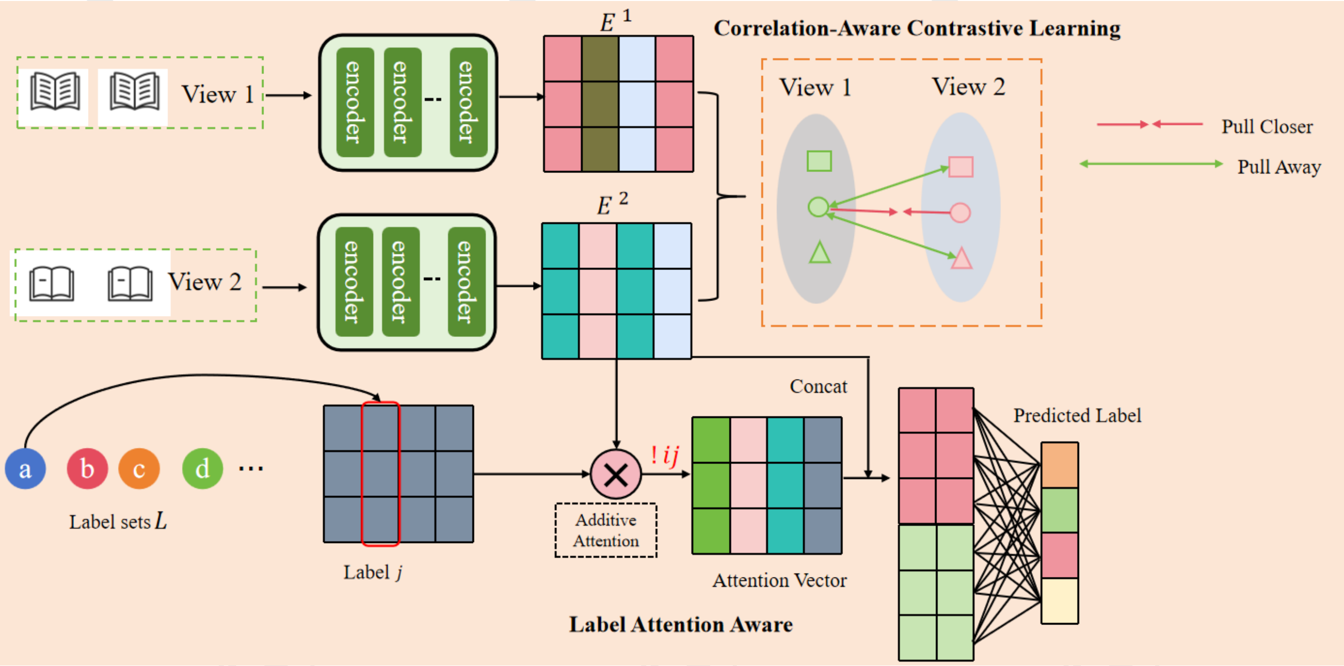


Figure 1: The overall framework of UCAL is as follows: First, data augmentation is applied to the text to obtain different views, and embeddings E^1 and E^2 for each view are generated through the encoder. Then, the model passes through the Correlation Aware Contrastive Learning module and the Label Attention Aware Module. The Label Attention Aware Module captures label dependencies by embedding labels and input text into a shared semantic space, effectively aligning text representations with label semantics. On the other hand, Correlation Aware Contrastive Learning dynamically models sample-level relationships to further refine these representations.

4 Method

4.1 Label Attention Aware Network

To capture the dependencies between label embeddings and text features, the text data is first embedded into a vector space. Let the text embeddings be denoted as $F = \{f_1, \ldots, f_N\}$, where $f_i \in \mathbb{R}^d$ represents the embedding of the i-th text instance. The interaction between a label embedding l_k and a text feature f_i is quantified using an additive attention mechanism, producing a score s_{ik}:

$$s_{ik} = \mathbf{v}_b^\top \tanh(\mathbf{W}_b[l_k; f_i]) \tag{2}$$

where $[a; b]$ denotes the concatenation of vectors a and b, and $\mathbf{v}_b$ and $\mathbf{W}_b$ are trainable parameters of the attention mechanism. The computed scores s_{ik} are then normalized across all text instances using the SoftMax function to obtain attention weights β_{ik}:

$$\beta_{ik} = \frac{\exp(s_{ik})}{\sum_{i'=1}^{N} \exp(s_{i'k})} \tag{3}$$

These attention weights are subsequently used to generate a text-aware attention vector $c_k \in \mathbb{R}^d$ by aggregating the text embeddings:

$$c_k = \sum_{i=1}^{N} \beta_{ik} f_i \tag{4}$$

The attention vector c_k encapsulates context-specific information from the text embeddings relevant to label l_k. To model the joint relationship between text and label, the attention vector c_k is concatenated with the label embedding l_k as $[l_k; c_k]$. This concatenated representation is passed through two fully connected layers, which refine the features and compute the final output[Zhou *et al.*, 2021]. The output scores for each text-label pair are transformed into probabilities using a sigmoid activation function, enabling multi-label predictions. By explicitly modeling text-aware label dependencies, this mechanism enhances the representation learning process, leading to improved performance in multi-label classification tasks.

5 Correlation Aware Contrastive Learning

In Multi-Label Text Classification (MLTC), identifying positive samples for the anchor point is complex due to the presence of multiple labels, unlike single-label classification where positive samples are clearly identified through exact label matching. Inspired by these preliminary insights[Zhang and Wu, 2024; Khosla *et al.*, 2020], we propose the Correlation Aware Contrastive Learning (CACL), which optimizes the contrastive learning process for both ”ANY” (partial label overlap) and ”ALL” (exact label match) scenarios. The two scenarios for defining positive samples are: For the general case, we denote $A(i)$ as the index of all samples involved in the loss calculation from both the batch and the queue. The supervised contrastive loss is defined as:

$$\mathcal{L}^{(i)}_{\text{supcon}} = -\frac{1}{|P^{(i)}|} \sum_{p \in P^{(i)}} \log \frac{\exp(s_p^{(i)}/\tau)}{\sum_{a \in A(i)} \exp(s_a^{(i)}/\tau)} \tag{5}$$

where $s_p^{(i)} = z_q^{(i)} \cdot z_k^{(p)}$, τ is a temperature scaling parameter, and $z_q^{(i)}$ and $z_k^{(p)}$ represent the L2-normalized query and key embeddings, respectively. By taking the gradient of $\mathcal{L}_{\text{supcon}}^{(i)}$ with respect to $z_q^{(i)}$, we calculate:

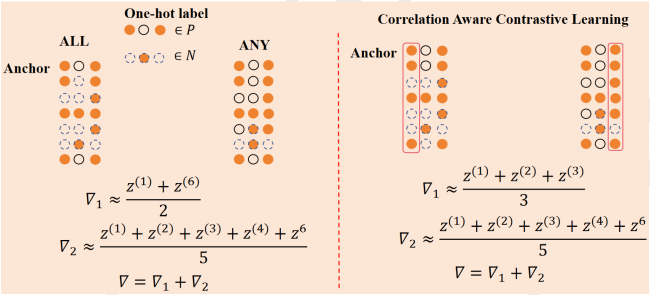

One-hot label
$\in P$
$\in N$
ALL
ANY
Anchor
$\nabla_1 \approx \frac{z^{(1)} + z^{(6)}}{2}$
$\nabla_2 \approx \frac{z^{(1)} + z^{(2)} + z^{(3)} + z^{(4)} + z^{6}}{5}$
$\nabla = \nabla_1 + \nabla_2$
Correlation Aware Contrastive Learning
Anchor
$\nabla_1 \approx \frac{z^{(1)} + z^{(2)} + z^{(3)}}{3}$
$\nabla_2 \approx \frac{z^{(1)} + z^{(2)} + z^{(3)} + z^{(4)} + z^{6}}{5}$
$\nabla = \nabla_1 + \nabla_2$

Figure 2: Illustration for **ALL** and **ANY** and method. Each row represents one sample's label, where the first row is the anchor and the following ones are samples in A, here $|A| = 6$. Each sample's label is denoted in a one-hot form where the light yellow circle means 1. The row with circles plotted with dotted lines means that the corresponding sample is in the negative set N, otherwise the sample is in the positive set P. The proposed Correlation Aware Contrastive loss function considers each label separately and forms multiple positive sets for one anchor sample. The positive sets for the anchor are $P_1 = \{1, 3, 6\}$, $P_3 = \{1, 2, 3, 4, 6\}$, suppose the anchor's label is $y = \{1, 3\}$.

$$\nabla_{z_q^{(i)}} \mathcal{L}_{\text{supcon}}^{(i)} = \sum_{p \in P^{(i)}} \frac{-1}{\tau |P^{(i)}|} z_k^{(p)} + \sum_{a \in A(i)} \frac{\exp(s_a^{(i)}/\tau)}{\sum_{b \in A(i)} \exp(s_b^{(i)}/\tau)} z_k^{(a)} \tag{6}$$

This gradient can be decomposed into two components:

$$\nabla_{z_q^{(i)}} \mathcal{L}_{\text{supcon}}^{(i)} = \bar{z} + \hat{z} \tag{7}$$

where:

$$\bar{z} = \frac{-1}{\tau |P^{(i)}|} \sum_{p \in P^{(i)}} z_k^{(p)} \tag{8}$$

$$\hat{z} = \sum_{a \in A(i)} \frac{\exp(s_a^{(i)}/\tau)}{\sum_{b \in A(i)} \exp(s_b^{(i)}/\tau)} z_k^{(a)} \tag{9}$$

Here, $\bar{z}$ represents the mean of the positive sample embeddings, while $\hat{z}$ represents the weighted average of all embeddings, which prevents representation collapse. This ensures the optimization direction moves $z_q^{(i)}$ towards a balanced representation of positive samples while considering all contextual dependencies. To further enhance the method, the proposed CACL loss treats the i-th sample as a distinct instance for each label $y_j^{(i)}$ it belongs to. A separate positive sample set is constructed for each $y_j^{(i)} \in y^{(i)}$ as:

$$P_j^{(i)} = \{m \mid y_j^{(i)} \in y^{(m)}\} \tag{10}$$

The CACL loss function is defined as:

$$\mathcal{L}_{\text{CACL}}^{(i)} = \sum_{y_j^{(i)} \in y^{(i)}} -\frac{1}{|P_j^{(i)}|} \sum_{p \in P_j^{(i)}} \log \frac{\exp(s_p^{(i)}/\tau)}{\sum_{a \in A(i)} \exp(s_a^{(i)}/\tau)} \tag{11}$$

This loss function generalizes the SupCon loss for multi-label tasks, reducing to the traditional SupCon loss in single-label scenarios where $|y^{(i)}| = 1$. Figure 2 illustrates how CACL effectively balances contributions from partially overlapping ("ANY") and fully matching ("ALL") scenarios, ensuring robust optimization for multi-label classification tasks.

- **ALL:** Only samples with exactly the same label set as the anchor are considered positive:
  $$P^{(i)} = \{m \mid y^{(m)} = y^{(i)}\}.$$
- **ANY:** Samples with any overlapping label class with the anchor are considered positive:
  $$P^{(i)} = \{m \mid y^{(m)} \cap y^{(i)} \neq \emptyset\}.$$

5.1 D. Training Objective

The UCLAF is an end-to-end MLTC model that is made possible by the constructed label attention and CACL. The goal of UCLAF is to minimize the target loss function L_{sum}, which includes L_{CACL} and $L_{\text{BCE}}(\hat{y}_{\text{doc}}^i, y^i)$. The specific formula is defined as:

$$L_{\text{sum}} = \lambda L_{\text{CACL}} + L_{\text{BCE}}(\hat{y}_{\text{text}}^i, y^i) \tag{12}$$

Here, λ is the tunable coefficient for the document-label contrastive learning loss function and the Correlation Aware Contrastive Learning (CACL) loss function, respectively, controlling the balance between the two losses. $\hat{y}_{\text{text}}^i$ represents the final probability that the semantic information of the i-th document is used as input.

6 IV. Experiments

In this section, we conduct a series of experiments on real-world text datasets to evaluate the performance of UCLAF and answer the following research questions:

Q1: Does UCLAF outperform current state-of-the-art (SOTA) methods? Q2: Does each module in UCLAF contribute to the overall model performance? Q3: How does correlation-aware contrastive learning operate and improve performance? Q4: How does the label attention network capture the relationships between labels and text? Q5: What is the impact of hyperparameters α and λ on UCLAF's performance?

6.1 Datasets

We evaluated our proposed model on two benchmark datasets: AAPD[Yang *et al.*, 2018] and EURLex[Loza Mencía and Fürnkranz, 2008]. Table 2 presents the statistics for these datasets, where N_{trn} represents the number of samples in the training set, L denotes the total number of labels, $\bar{L}$ is the average number of labels per document, and $\bar{W}_{\text{trn}}$ represents the average number of words per training document.

6.2 Baseline Methods and Evaluation Metrics

To rigorously assess the effectiveness of the proposed UCLAF method, we compare it against several state-of-the-art multi-label text classification (MLTC) models that have been widely recognized in recent research. The following baseline methods are selected for comparison:

- **XML-CNN**[Liu *et al.*, 2017b]: A CNN-based model that employs a dynamic max pooling scheme to capture high-level features in the input text.
- **AttentionXML**[You *et al.*, 2019]: A BiLSTM-based model that integrates a label tree structure, utilizing label attention to capture contextual semantics and generate label-aware document representations.
- **CornetAttentionXML**[Xun *et al.*, 2020]: An architecture that combines the AttentionXML text encoder with a Cornet module to effectively model label correlations and dependencies in multi-label classification.
- **LightXML**[Jiang *et al.*, 2021]: A lightweight deep learning framework that incorporates dynamic negative label sampling. For experimental consistency, we reproduce this model using BERT as the underlying architecture.
- **GUDN**[Wang *et al.*, 2023]: A model that utilizes guided networks to enhance classification performance by improving the representation of labels.
- **MLGN**[Liu *et al.*, 2023]: A method that integrates label semantic information to improve document representations but does not fully address complex label dependencies in multi-label tasks, nor does it effectively model label associations.

6.3 Evaluation Metrics

We evaluate model performance using two widely adopted metrics: precision at k (P@k) and normalized discounted cumulative gain at k (N@k)[You *et al.*, 2019; Xiao *et al.*, 2019],with higher values indicating better performance.

- **P@k**: Measures the proportion of relevant items among the top-k results.
- **N@k**: Considers both the relevance and ranking of retrieved items, rewarding relevant items ranked higher.

Datasets	N_{trn}	N_{test}	D	L	$\bar{L}$	$\bar{W}_{trn}$	$\bar{W}_{test}$
EURLex	15499	3865	186104	3956	5.30	1248	1230
AAPD	54840	1000	69399	54	2.41	163	171

Table 2: Data statistics.

6.4 Implementation Details

All experiments were conducted on an NVIDIA 3090 GPU using the pre-trained bert-base-uncased model for feature extraction. The maximum input text length was set to 512, with a dropout rate of 0.5 for text representation, and a batch size of 16. For the EURLex dataset, the learning rate was set to 5×10^{-5}, with $\lambda = 0.5$ and $\tau = 3$. For the AAPD dataset, the learning rate was set to 1×10^{-5}, with $\lambda = 0.7$ and $\tau = 5$.

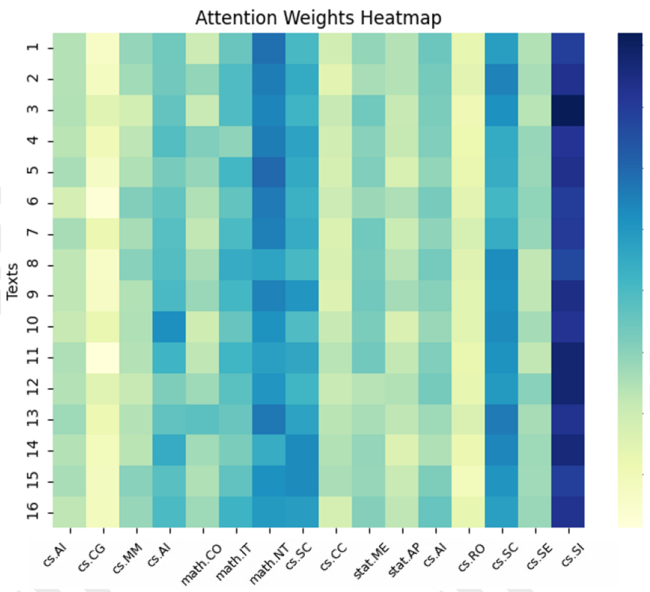


Figure 3: Attention map on AAPD. Due to space limitation, we only report the results for the first 16 label.

6.5 Main Results(Q1)

We evaluated the performance of the proposed UCLAF model against state-of-the-art methods using P@K and N@K as evaluation metrics. The results in Table 3 show that UCLAF achieves superior performance compared to existing models across all metrics. Specifically, UCLAF outperforms the second-best model in terms of P@1, P@3, P@5, N@3, and N@5 by significant margins. For the EURLex dataset, UCLAF improves P@1 by 1.01%, P@3 by 2.62%, P@5 by 0.24%, N@3 by 0.69%, and N@5 by 0.40% compared to the second-best model, which is AttentionXML. Similarly, for the AAPD dataset, UCLAF improves P@1 by 0.59%, P@3 by 0.39%, P@5 by 0.28%, N@3 by 0.54%, and N@5 by 0.53% compared to the second-best model, which is MLGN. These results demonstrate that UCLAF consistently leads to better classification performance, confirming its effectiveness in multi-label text classification tasks.

7 ANALYSIS

7.1 Ablation Study (Q2)

We conducted ablation experiments to evaluate the impact of the Label Attention Aware Network (LAN) and Correlation-Aware Contrastive Learning (CACL) modules in UCLAF, with results shown in Table 4. Adding the LAN module improves performance by leveraging label semantics. For example, P@1 increases from 85.30 to 85.58 on AAPD and from 84.55 to 85.42 on EURLex. Similarly, incorporating the CACL module enhances performance by capturing label correlations, as seen in the rise of P@5 from 41.18 to 42.24 on AAPD and P@3 from 73.05 to 74.13 on EURLex. The combination of both LAN and CACL achieves the best results across all metrics, with P@1 reaching 86.69 on AAPD and 87.32 on EURLex. These results demonstrate that LAN

Datasets		XMLCNN	AttentionXML	CornetAttentionXML	LightXML	CUDN	MLGN	UCLAF
EURLex[a]	P@1	76.81	85.90	85.85	86.03	85.51	86.31	**87.32**
	P@3	62.79	73.01	73.32	74.19	74.10	74.77	**75.39**
	P@5	51.56	61.00	61.68	62.27	62.14	62.66	**62.90**
	N@3	66.44	76.41	76.61	77.41	77.23	77.97	**78.66**
	N@5	60.47	70.47	70.49	71.70	71.49	72.13	**72.50**
AAPD[b]	P@1	74.38	83.70	86.00	85.80	85.80	86.10	**86.69**
	P@3	53.84	60.63	61.57	61.30	62.30	62.57	**62.96**
	P@5	37.79	41,64	41.76	42.02	42.42	42.44	**42.72**
	N@3	71.12	79.90	81.25	81.26	81.97	82.45	**82.99**
	N@5	75.93	84.10	84.90	85.32	85.87	86.17	**86.70**

Table 3: Comparisons with state-of-the-art and representative methods.

Method			AAPD					EURLex				
Baseline	LAN	CACL	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
Y	N	N	85.30	61.37	41.18	82.45	81.30	84.55	73.05	61.49	80.93	79.20
Y	Y	N	85.58	61.67	41.26	82.87	81.85	85.42	73.87	62.18	81.65	79.86
Y	N	Y	85.78	61.97	42.24	82.21	82.10	86.20	74.13	62.37	77.82	71.67
Y	Y	Y	**86.69**	**62.96**	**42.72**	**82.99**	**86.70**	**87.32**	**75.39**	**62.90**	**78.66**	**72.50**

Table 4: Comparison of the ablation results of each module.

and CACL complement each other effectively, jointly contributing to significant improvements in label alignment and inter-label modeling.

7.2 Analysis of Label Attention Aware (Q3)

The heatmap in AAPD dataset demonstrates the Label Attention Module's ability to capture semantic relationships between texts and labels by aligning them in a shared semantic space. For example, *Label 1* shows strong attention on *Text 3* and *Text 5*, indicating effective focus on relevant text segments. Overlapping attention patterns, such as for *Label 2* and *Label 3* on *Text 4* and *Text 6*, highlight the module's ability to model label correlations, capturing co-occurrence and semantic dependencies between labels.

However, limitations exist. For example, *Text 7* consistently receives low attention weights, suggesting weak alignment or poor feature representation. Similarly, *Label 4* shows a uniform attention distribution, reducing precision in distinguishing relationships. These results confirm the module's effectiveness in addressing complex label correlations while highlighting areas for further improvement.

7.3 Correlation Aware Contrastive Learning Experiments(Q4)

To evaluate the effectiveness of CACL, we conducted a series of experiments. First, we examined the impact of positive and negative sample pair construction on model performance. Second, we visualized the learned sample representations to provide qualitative insights. Finally, we investigated the influence of different data augmentation strategies on the model's overall performance.

Effects of Positive Sample Generation

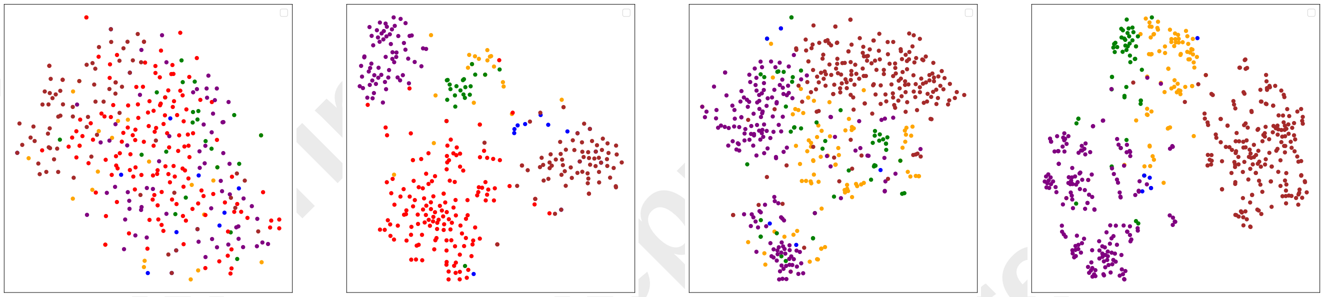
As shown in Table 5, CACL significantly outperforms the ”ANY” and ”ALL” match strategies on the AAPD dataset. It achieves the highest scores across all metrics, with a P@1 of 87.32, compared to 85.46 (ANY) and 84.32 (ALL). Similar improvements are observed in P@3 and P@5. The key strength of CACL lies in its flexible positive sample selection, associating samples with others that share a subset of their labels. This strategy better suits multi-label tasks by capturing label correlations, leading to more effective representation learning and performance gains.

Performance of Different Data Augmentation Techniques

Table 6 presents a comparison of different data augmentation strategies, including random masking, continuous masking, token shuffling, and the dropout strategy. Among these, the dropout strategy achieves the highest performance, indicating its effectiveness in generating positive samples by randomly removing neuron-level information. Random masking and continuous masking yield comparable results, showing no significant differences in their impact on model performance. Conversely, token shuffling results in the lowest scores, likely due to its disruption of text semantics, which hinders the model's ability to recover meaningful information. These findings underscore the critical role of maintaining semantic structure in data augmentation to ensure effective learning.

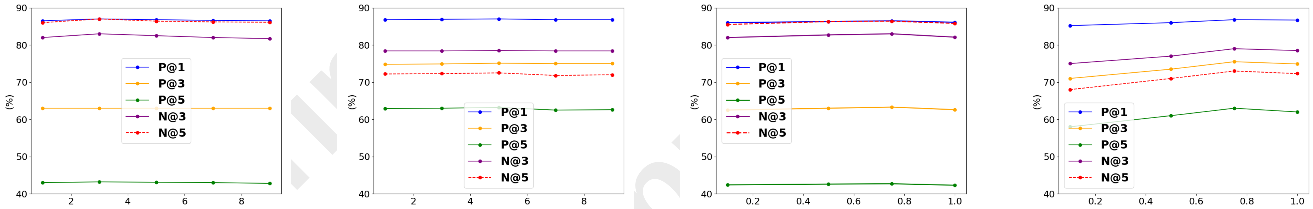
Visualization of Learned Representation

To further investigate the interpretability of the high-quality document representations in UCLAF, this section provides a visual analysis of the document representations in UCLAF. We employ t-SNE to visualize the category prototypes, as shown in Figure 4. The results demonstrate that the category prototypes capture meaningful semantic information. Specifically, the distance between different categories is maximized, while the distance between pairs of highly co-occurring categories is minimized, and pairs of categories with low co-occurrence are more distantly separated.



(a) BCE(AAPD) (b) Ours(AAPD) (c) BCE(EURLex) (d) Ours(EURLex)

Figure 4: Visualization of Learned Representation.



(a) AAPD (b) EURLex (c) AAPD (d) EURLex

Figure 5: Hyper-parameter Analysis. Figures (a) and (b) show temperature parameters (τ), and (c) and (d) show balancing parameters (λ).

Strategy	P@1	P@3	P@5
ALL	84.32	73.59	60.34
ANY	85.46	74.18	60.62
CACL	**87.32**	**75.39**	**62.90**

Table 5: Performance of different contrastive losses on the AAPD dataset.

Strategy	P@1	P@3	P@5
Random Masking	87.15	75.27	62.36
Continuous Masking	86.92	74.56	61.85
Token Shuffling	86.61	74.21	61.72
Ours	**87.32**	**75.39**	**62.90**

Table 6: Performance of different positive example generation techniques on the AAPD dataset.

7.4 Hyper-parameter Analysis(Q5)

Sensitivity Analysis of hyper-parameter threshold λ

λ controls the balance between the contrastive learning loss L_{CACL} and the binary cross-entropy loss L_{BCE}. As is shown in Figure 5(a)(b), when λ is too large, the model overly emphasizes label correlations, which may reduce the performance of individual label classification. Conversely, when λ is too small, the model focuses more on independent label classification but may overlook global semantic relationships between labels. Therefore, λ needs to be adjusted experimentally to balance label correlation modeling and text-label matching performance.

Sensitivity analysis of the hyperparameter τ

The parameter τ controls the sharpness of the similarity distribution in contrastive learning. As is shown in Figure 5(c)(d), when τ is too large, the similarity scores become too uniform, reducing the model's ability to distinguish between positive and negative samples, which weakens its performance. Conversely, when τ is too small, the similarity distribution becomes overly sharp, causing the model to overfit a few high-similarity positive samples while ignoring others, leading to poor generalization. Therefore, τ needs to be carefully tuned to balance the focus on high-quality positive samples and the overall similarity distribution, typically through validation experiments.

8 Conclusion

We proposed the Unified Contextual and Label-Aware Framework (UCLAF) to tackle challenges in multi-label text classification, such as label interdependencies and overlaps. By combining label attention and supervised contrastive learning, UCLAF captures semantic relationships and reduces label ambiguity. Experiments on AAPD, EURLex show that UCLAF outperforms state-of-the-art methods . However, UCLAF relies on static label embeddings, struggles with large-scale datasets due to contrastive learning's computational cost, and is limited to text data. Future work could explore dynamic label embeddings, improve efficiency, and extend UCLAF to multi-modal or hierarchical tasks for greater scalability and versatility.

References

[Bai *et al.*, 2020] Junwen Bai, Shufeng Kong, and Carla Gomes. Disentangled variational autoencoder based multi-label classification with covariance-aware multivariate probit model. *arXiv preprint arXiv:2007.06126*, 2020.

[Bai *et al.*, 2022] Junwen Bai, Shufeng Kong, and Carla P Gomes. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *international conference on machine learning*, pages 1383–1398. PMLR, 2022.

[Bogatinovski *et al.*, 2022] Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215, 2022.

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[Cheng *et al.*, 2025] Yuqi Cheng, Yunkang Cao, Dongfang Wang, Weiming Shen, and Wenlong Li. Boosting global-local feature matching via anomaly synthesis for multi-class point cloud anomaly detection. *IEEE Transactions on Automation Science and Engineering*, 2025.

[Durand *et al.*, 2019] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[Ge *et al.*, 2018] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1277–1286, 2018.

[Guo *et al.*, 2021] Biyang Guo, Songqiao Han, Xiao Han, Hailiang Huang, and Ting Lu. Label confusion learning to enhance text classification models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12929–12936, 2021.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[Jiang *et al.*, 2021] Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7987–7994, 2021.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[Kurata *et al.*, 2016] Gakuto Kurata, Bing Xiang, and Bowen Zhou. Improved neural network-based multi-label classification with better initialization leveraging label cooccurrence. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–526, 2016.

[Liu *et al.*, 2016] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

[Liu *et al.*, 2017a] 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*, pages 115–124, 2017.

[Liu *et al.*, 2017b] 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*, pages 115–124, 2017.

[Liu *et al.*, 2021] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7955–7974, 2021.

[Liu *et al.*, 2023] Qiang Liu, Jingzhe Chen, Fan Chen, Kejie Fang, Peng An, Yiming Zhang, and Shiyu Du. Mlgn: A multi-label guided network for improving text classification. *IEEE Access*, 2023.

[Loza Mencía and Fürnkranz, 2008] Eneldo Loza Mencía and Johannes Fürnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 50–65. Springer, 2008.

[Ma *et al.*, 2021] Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 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, 2021.

[Moschitti *et al.*, 2014] Alessandro Moschitti, Bo Pang, and Walter Daelemans. Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[Nam *et al.*, 2017] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. *Advances in neural information processing systems*, 30, 2017.

[Read *et al.*, 2011] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

[Rubin *et al.*, 2012] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88:157–208, 2012.

[Wang *et al.*, 2016] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.

[Wang *et al.*, 2018] Kaixiang Wang, Ming Yang, Wanqi Yang, and YiLong Yin. Deep correlation structure preserved label space embedding for multi-label classification. In *Asian Conference on Machine Learning*, pages 1–16. PMLR, 2018.

[Wang *et al.*, 2021] Ran Wang, Robert Ridley, Weiguang Qu, Xinyu Dai, et al. A novel reasoning mechanism for multi-label text classification. *Information Processing & Management*, 58(2):102441, 2021.

[Wang *et al.*, 2022] Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. *arXiv preprint arXiv:2203.03825*, 2022.

[Wang *et al.*, 2023] Qing Wang, Jia Zhu, Hongji Shu, Kwame Omono Asamoah, Jianyang Shi, and Cong Zhou. Gudn: A novel guide network with label reinforcement strategy for extreme multi-label text classification. *Journal of King Saud University-Computer and Information Sciences*, 35(4):161–171, 2023.

[Wu *et al.*, 2020] Jian Wu, Victor S Sheng, Jing Zhang, Hua Li, Tetiana Dadakova, Christine Leon Swisher, Zhiming Cui, and Pengpeng Zhao. Multi-label active learning algorithms for image classification: Overview and future promise. *ACM Computing Surveys (CSUR)*, 53(2):1–35, 2020.

[Xiao *et al.*, 2019] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. Label-specific document representation for multi-label text classification. 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 466–475, 2019.

[Xiao *et al.*, 2021] Yaoqiang Xiao, Yi Li, Jin Yuan, Songrui Guo, Yi Xiao, and Zhiyong Li. History-based attention in seq2seq model for multi-label text classification. *Knowledge-Based Systems*, 224:107094, 2021.

[Xu *et al.*, 2023] Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. Label-specific feature augmentation for long-tailed multi-label text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 10602–10610, 2023.

[Xun *et al.*, 2020] Guangxu Xun, Kishlay Jha, Jianhui Sun, and Aidong Zhang. Correlation networks for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1074–1082, 2020.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[Yang *et al.*, 2018] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. Sgm: sequence generation model for multi-label classification. *arXiv preprint arXiv:1806.04822*, 2018.

[Yazici *et al.*, 2020] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020.

[Yeh *et al.*, 2017] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent space for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Ying *et al.*, 2021] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

[You *et al.*, 2019] 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. *Advances in neural information processing systems*, 32, 2019.

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

[Zhang *et al.*, 2019] Suwei Zhang, Yuan Yao, Feng Xu, Hanghang Tong, Xiaohui Yan, and Jian Lu. Hashtag recommendation for photo sharing services. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5805–5812, 2019.

[Zhao *et al.*, 2021] Wenting Zhao, Shufeng Kong, Junwen Bai, Daniel Fink, and Carla Gomes. Hot-vae: Learning high-order label correlation for multi-label classification via attention-based variational autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 15016–15024, 2021.

[Zhou *et al.*, 2021] Cangqi Zhou, Hui Chen, Jing Zhang, Qianmu Li, Dianming Hu, and Victor S Sheng. Multi-label graph node classification with label attentive neighborhood convolution. *Expert Systems with Applications*, 180:115063, 2021.