External Knowledge-Enhanced Semi-Supervised Multi-Label Short Text Classification

Zhi-Jie Wang[1], Yirui Li[1], Shuhui Cao[1], and Peipei Li[2](✉)

[1] College of Computer Science, Chongqing University, Chongqing 400044, China
[2] School of Computer Science and Information Engineering, Hefei University of Technology, Anhui 230009, China
peipeili@hfut.edu.cn

Abstract. Multi-label Short Text Classification (MSTC) focuses on assigning multiple relevant labels to each short text. However, it is confronted with several challenges such as data sparsity, label scarcity, and class imbalance. Motivated by this, in this paper, we introduce a novel MSTC method called Knowledge-based Long-tail Data Augmentation and Pseudo-label optimization (KLDAP). Specifically, to fix the data sparsity, KLDAP firstly enriches the feature representation of short texts by integrating external knowledge from concept knowledge graphs and entity recognition techniques. Secondly, it incorporates an optimized semi-supervised learning framework to enhance the utilization of unlabeled data with the help of external knowledge and generated pseudo labels. Thirdly, to further address the class imbalance, we introduce an innovative data augmentation strategy that combines a Variational Autoencoder (VAE) with head label feature transfer and contrastive learning, optimizing the representation of tail labels. Finally, extensive experiments conducted on four benchmark datasets under varying labeling ratios demonstrate that KLDAP significantly outperforms state-of-the-art methods, effectively tackling the challenges in MSTC.

Keywords: Multi-label Short Text Classification · Semi-supervised Learning · Contrastive Learning

1 Introduction

With the advancement of internet technology and the widespread usage of mobile devices, short text data has become a key medium for sharing information [1]. The rapid growth of short texts on social media, messaging apps, and online reviews creates both opportunities for analysis and technical challenges [2]. Multi-label Short Text Classification (MSTC), which assigns multiple labels to each short text, is applied in areas like social media analysis [3], news aggregation [4], and e-commerce recommendations [5], but its development still faces the following several challenges.

Firstly, as compared to long texts, short texts face significant data sparsity due to limited context, often resulting in unsatisfactory multi-label classification results. To address

D. Huang et al. (Eds.): ICIC 2025, LNAI 15864, pp. 65–76, 2025.
https://doi.org/10.1007/978-981-95-0014-7_6

this, researchers either optimize short texts' internal features by adjusting model architectures [6] or enhance their semantic representation by integrating external knowledge like word concepts and topics.

In addition to semantic sparsity, the lack of labeled data further constrains MSTC development. Acquiring large-scale, high-quality labeled data in real-world scenarios requires significant resources, making it difficult for models to learn effective label–text relationships. This scarcity often leads to overfitting and weak generalization ability. Some researchers have explored data augmentation methods, such as randomly replacing, or deleting words to expand the training dataset [7]. However, these methods will introduce noise and disrupt the semantic structure, negatively impacting classification performance. In contrast, Semi-Supervised Learning (SSL) leverages unlabeled data to enhance model training.

Moreover, the common phenomenon of class imbalance in MSTC further intensifies the challenge of classification. For example, in Tweet dataset, a few labels like "joy" and "disgust" have over 4000 samples (head labels), while the majority of labels, such as "trust" and "surprise", appear only rarely (tail labels), forming a long-tail distribution. This imbalance leads the classifier to favor head labels, while suffering from the poor performance on tail labels. At the meanwhile, insufficient labeled data further hinders the learning of tail label features, worsening the imbalance.

To overcome the above challenges, this paper proposes a Knowledge-based Long-tail Data Augmentation and Pseudo-label optimization method (KLDAP) for MSTC. Specifically, KLDAP enhances short text features by integrating external knowledge, while an optimized semi-supervised learning framework leverages unlabeled data through dynamic pseudo-label refinement to boost generalization. Additionally, an innovative data augmentation strategy combining a Variational Autoencoder (VAE) with head label feature transfer and contrastive learning enhances tail label representations and mitigates label imbalance. Experimental results demonstrate that KLDAP significantly improves classification performance on multiple MSTC benchmark datasets.

The main contributions of this paper are:

1. A new framework KLDAP is proposed to address key challenges in multi-label short text classification, including data sparsity, label scarcity, and long-tail imbalance;
2. An optimized pseudo-label generation method is developed to enhance unlabeled data utilization;
3. A VAE-based data augmentation strategy with head label transfer and contrastive learning is introduced for tail-label representation;
4. Extensive experiments demonstrate that KLDAP outperforms previous SOTA methods on four benchmark datasets.

2 Related Works

Multi-label Short Text Classification has attracted significant attention as an extension of traditional text classification [8]. Existing solutions are typically divided into problem transformation methods and algorithm adaptation methods. The former converts multi-label tasks into single-label problems but incurs high costs and ignores label correlations.

The latter modifies single-label algorithms for better efficiency but captures only low-order correlations. With the advancement of deep learning technologies, MSTC techniques based on neural networks have seen significant growth. Early CNN-based models [9] have improved performance, but they still suffer from data sparsity. To mitigate this, external knowledge-based approaches have been employed to expand the feature space, though their effectiveness depends heavily on data quality. Meanwhile, the Transformer-based models, notably BERT [10], have further advanced MSTC. Researchers such as Yarullin et al. [11] and Chen et al. [12] have integrated external knowledge with BERT to address semantic sparsity, but these methods often require large amounts of labeled data.

Semi-supervised learning bridges supervised and unsupervised learning [13], utilizing limited labeled data and abundant unlabeled data. Methods such as consistency regularization and pseudo-labeling have been widely applied to MSTC [14]. The former enhances model robustness, while the latter improves generalization by leveraging unlabeled data. For example, Wang et al. [15] introduced a self-tuning framework with Pseudo-Group Contrast (PGC) to mitigate model shift and pseudo-label bias. Yang et al. [16] proposed a prototype-guided approach to refine decision boundaries and address data imbalance.

In summary, significant research achievements have been made in MSTC and semi-supervised learning. However, as far as we know, there are no related works on semi-supervised MSTC. The KLDAP method proposed aims to comprehensively address the shortcomings of existing research by incorporating external knowledge, optimizing the SSL framework, and innovating data augmentation strategies.

3 Methodology

3.1 Problem Definition

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be the training set consisting of N short text samples, where each x_i represents a short text, and $y_i \in Y = \{0, 1\}^L$ is the label vector associated with x_i, with L denoting the total number of labels. Each label $y_{i,j} \in \{0, 1\}$ indicates whether the short text x_i is associated with label j. The MSTC task in this study focuses on learning an optimal mapping from D to Y, enabling the model to assign the most appropriate labels to each test short text after training.

3.2 Overall Architecture

The model architecture of KLDAP is shown in Fig. 1, which mainly consists of three core modules: Fundamental MSTC Model, Semi-Supervised Module, and Data Augmentation Module. Initially, we use both the text and external knowledge from Probase[1] for concept retrieval. Retrieving the top K related concepts expands the word space. In terms of the extended feature space, the fundamental MSTC model is initially built. In the semi-supervised Module, a pseudo-label selector accurately selects positive and negative samples and generates pseudo-labels that are merged with the original training

[1] https://probasegroup.com/.

set to enhance generalization. Finally, the data augmentation module uses a VAE model with head label feature transfer and contrastive learning to optimize tail label features, increasing robustness against rare labels. Next, we will provide a detailed introduction to the key modules of the proposed method and their working principles.

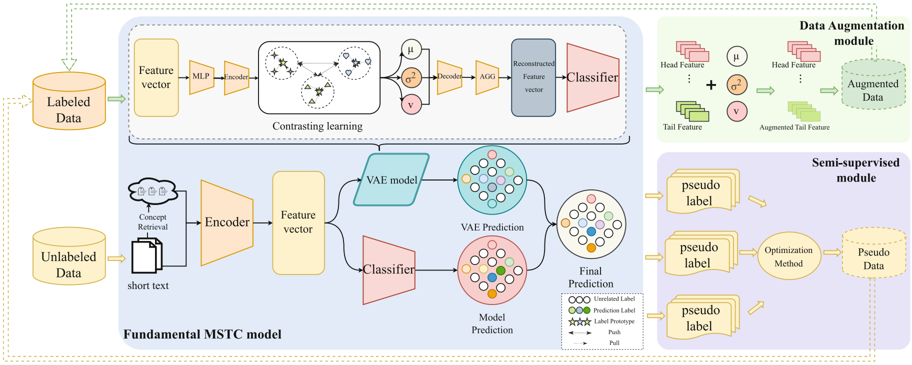

Fig. 1. The architecture of our KLDAP model.

3.3 Concept Retrieval Module

In MSTC tasks, short texts often lack sufficient context due to their brevity. To enrich their sparse semantic information, we extend their representation using external conceptual knowledge from Probase, an open knowledge graph by Microsoft that provides probabilistic concept associations. During expansion, only highly relevant concepts are retained. As shown in Fig. 2, This module ensures effective semantic enhancement.

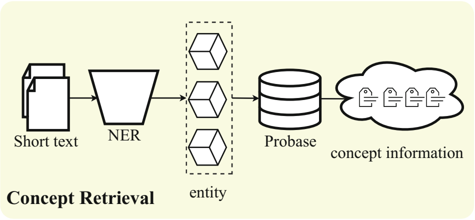

Fig. 2. The architecture of the Concept Retrieval module.

To extract entities from short texts, we use Named Entity Recognition (NER). For instance, from the sentence “Apple announced the launch of a new iPhone model.” NER extracts “Apple” and “iPhone model.” Since these entities might not match the standard forms in the knowledge base, we employ the Tagme[2] tool to normalize them and preprocess abbreviations to improve matching accuracy. Finally, we retrieve relevant concept

[2] https://sobigdata.d4science.org/web/tagme/tagme-help.

information from Probase, effectively incorporating external knowledge to strengthen semantic representation and mitigate text sparsity.

3.4 Fundamental MSTC Model

The proposed MSTC model consists of two main components: (i) the feature extraction module based on the Pre-trained Language Model (PLM) and (ii) the label optimization module based on the VAE model and contrastive learning strategy.

Feature Extraction Module. For each short text x_i, external conceptual knowledge c^i_{ext} is first extracted to enrich its semantics. We then fine-tune a pre-trained language model (PLM) by concatenating x_i and c^i_{ext} to obtain the semantic feature:

$$\mathbf{h}_i = \text{PLM}\left(x_i \oplus c^i_{\text{ext}}; \theta_{\text{PLM}}\right) \tag{1}$$

where $\oplus$ denotes concatenation, θ_{PLM} denotes the parameters of the PLM.

Next, the feature vector $\mathbf{h}_i$ is input into a multi-label classifier f_{cls}, which generates the predicted value $\hat{y}^{\text{PLM}}_{i,j} = f_{\text{cls}}(\mathbf{h}_i)$, where $\hat{y}^{\text{PLM}}_{i,j}$ is the predicted probability for the j-th label of i-th sample. The prediction is optimized using a binary cross-entropy loss:

$$L_{\text{clf}} = -\sum_{j=1}^{L}\left[y_{i,j}\log\left(\hat{y}^{\text{PLM}}_{i,j}\right) + \left(1 - y_{i,j}\right)\log\left(1 - \hat{y}^{\text{PLM}}_{i,j}\right)\right] \tag{2}$$

where $y_{i,j} \in \{0, 1\}$ represents the true value of the j-th label for the i-th sample.

Label Optimization Module. In the MSTC task, texts are often annotated with multiple labels that exhibit overlapping as well as distinct semantic properties. Conventional sample-based methods may conflate shared and label-specific features, leading to ambiguous representations. To tackle this issue, we propose a label optimization module utilizing label-wise decoupling, a VAE model, and contrastive learning.

Specifically, given a global text feature $\mathbf{h}_i$, we employ a set of label-specific decoupling functions $\{g_j\}_{j=1}^{L}$ to decompose $\mathbf{h}_i$ into label-specific sub-features:

$$h_{i,j} = g_j(\mathbf{h}_i), \Delta j = 1, 2, \ldots, L \tag{3}$$

where $h_{i,j} \in \mathbb{R}^{d_j}$ represents the sub-feature corresponding to label $y_{i,j}$. Each sub-feature $h_{i,j}$ is then processed by a VAE model, which encodes it into a latent distribution $q_\phi\left(z_{i,j}|h_{i,j}\right)$ and decodes it to reconstruct $h_{i,j}$, thereby ensuring that the latent space retains the essential semantic information.

To further enforce label-wise independence and reduce inter-label interference, we integrate a contrastive learning mechanism. For a mini-batch of size B, all the latent embeddings can be aggregated into a set $\mathcal{Z}$ as follows:

$$\mathcal{Z} = \left\{z_{i,j} i = 1, \ldots, B, j = 1, \ldots, L\right\} \tag{4}$$

where $z_{i,j}$ represents the latent embedding corresponding to sample i and label j. This set $\mathcal{Z}$ forms the contrastive embedding pool $\mathcal{A}$. For each embedding $z_{i,j}$, we need to define its positive and negative sample sets based on the label information. The positive samples refer to embeddings that share the same label as the anchor $z_{i,j}$. Specifically, for a given embedding $z_{i,j}$, its positive sample set $P\left(z_{i,j}\right)$ consists of other sample embeddings that belong to the same label j:

$$P\left(z_{i,j}\right) = \left\{z_{i',j} | i' \neq i,\ y_{i',j} = 1\right\} \tag{5}$$

The negative samples refer to embeddings that correspond to different labels from the anchor $z_{i,j}$. For a given embedding $z_{i,j}$, its negative sample set $\mathcal{A}\left(z_{i,j}\right)$ consists of all embeddings with different labels:

$$\mathcal{A}\left(z_{i,j}\right) = Z \backslash \left\{z_{i,j}, P\left(z_{i,j}\right)\right\} \tag{6}$$

Then, we use contrastive loss to optimize the model, ensuring that positive samples are pulled closer, while negative samples are pushed further apart. The corresponding contrastive loss is formulated as:

$$L_{\text{psc}} = -\frac{1}{\left|P\left(z_{i,j}\right)\right|} \sum_{z^{+} \in P\left(z_{i,j}\right)} \log \frac{\exp\left(\frac{z_{i,j}^{\top} z^{+}}{\tau}\right)}{\sum_{z' \in \mathcal{A}\left(z_{i,j}\right)} \exp\left(\frac{z_{i,j}^{\top} z'}{\tau}\right)} \tag{7}$$

where τ is a temperature hyperparameter that scales the similarity scores.

Subsequently, each latent embedding $z_{i,j}$ is decoded via a shared decoder $d(\cdot)$ to yield the decoded representation $\tilde{u}_{i,j}$. For each sample x_i, the decoded vectors are aggregated using an aggregation function $a(\cdot)$ to form the reconstructed feature:

$$\tilde{\mathbf{h}}_i = a\left(\left[\tilde{u}_{i,1}, \tilde{u}_{i,2}, \ldots, \tilde{u}_{i,L}\right]\right) \tag{8}$$

To ensure that the reconstructed feature vector $\tilde{\mathbf{h}}_i$ preserves the original semantic content of the input text, we introduce a reconstruction loss based on the L2 norm:

$$L_{\text{rec}} = \|\mathbf{h}_i - \tilde{\mathbf{h}}_i\|^2 \tag{9}$$

Then, a classifier f_{VAE} predicts label-specific outputs $\hat{y}_{i,j}^{\text{VAE}}$ based on $\tilde{\mathbf{h}}_i$. The overall prediction integrates the PLM-based output and the label optimization prediction:

$$\hat{y}_{i,j} = \alpha\, \hat{y}_{i,j}^{\text{PLM}} + (1 - \alpha)\, \hat{y}_{i,j}^{\text{VAE}} \tag{10}$$

where α is a hyperparameter that determines the relative weight of each prediction.

For end-to-end model training, we define the total loss function:

$$L_{\text{total}} = L_{\text{BCE}} + \lambda_1 L_{\text{psc}} + \lambda_2 L_{\text{rec}} \tag{11}$$

where L_{BCE} denotes the binary cross entropy loss, and λ_1, λ_2 are hyperparameters balancing the contrastive and reconstruction losses. This comprehensive loss function enables the model to jointly optimize label-specific feature decoupling, semantic consistency, and classification accuracy.

3.5 Semi-Supervised Module

As mentioned earlier, the quality and quantity of labeled data are critical to the performance of classification model. However, labeled data is often scarce in MSTC task. To address this, we propose a pseudo-label-based semi-supervised learning framework.

Specifically, we define the labeled dataset as $D_L = \{(x_i, y_i)\}_{i=1}^{M}$, where M represents the number of labeled dataset. Similarly, the unlabeled dataset is defined as $D_U = \{x_i\}_{i=1}^{Q}$, containing Q samples. First, we train an initial model f on the labeled dataset D_L through supervised learning. Next, the model is used to predict the unlabeled dataset D_U under multiple random perturbations, calculating the mean prediction $\mathbb{E}\left[p_j(x)\right]$ and variance $\mathrm{Var}\left[p_j(x)\right]$ for each label j. The mean prediction is used to measure the model's confidence in the labels, while the variance serves as an uncertainty measure. These statistics provide a basis for selecting pseudo-labels. Subsequently, two key thresholds T_p and K_p are set to determine whether a label is assigned a positive pseudo-label. For each unlabeled sample x_i, the pseudo-label assignment is defined as follows:

$$c_{i,j} = \begin{cases} 1, \text{ if } \mathbb{E}\left[p_j(x_i)\right] \geq T_p \text{ and } \mathrm{Var}\left[p_j(x_i)\right] \leq K_p \\ 0, \text{ otherwise} \end{cases} \tag{12}$$

where $c_{i,j} = 1$ indicates a positive label and $c_{i,j} = 0$ indicates a negative label.

Finally, the pseudo-label vectors for all samples are combined with the labeled dataset D_L to construct a mixed dataset $D_{\text{mix}} = D_L \cup \left\{(x_i, \hat{y}_i) | c_{i,j} \neq 0\right\}$, where $\hat{y}_i$ is the multi-label vector generated by pseudo-labels.

3.6 Data Augmentation Module

To better tackle class imbalance, we introduce a data augmentation module that enhances feature diversity and generalization by expanding tail features.

After training the fundamental model, we obtain the latent feature space $\mathscr{X}$ and compute each prototype vector $p_j = \frac{1}{n_j} \sum_{i=1}^{n_j} z_{i,j}$ by averaging its latent features, where n_j denotes the total number of samples associated with label j. To distinguish between head labels and tail labels, we introduce a hyperparameter $N_t \in \mathbb{R}^+$ as a threshold. A label j is categorized as a tail label if its sample count n_j is less than N_t; otherwise, it is a head label. Consequently, we obtain the head label set H and the tail label set T.

To address tail label feature scarcity, we adopt a feature transfer method based on head labels. For a tail label $t \in T$, we randomly select a head label $h \in H$ and generate N_a augmented features for the tail label prototype:

$$z_t^k = p_t + \sigma_h \odot \epsilon \tag{13}$$

Then, we carefully select samples containing tail labels from the labeled dataset D_L and apply a feature replacement strategy:

$$\hat{z}_i = \begin{cases} \alpha z_{i,j} + \beta z_t^k, \text{ if } \ t \in T \ \text{ and } y_{i,j} = 1 \\ z_{i,j}, \qquad\qquad \text{otherwise} \end{cases} \tag{14}$$

where α and β are predefined weights that adjust the contribution of the original and augmented features in the final replacement. After applying the feature replacement, the augmented feature vector z_i is fed into the decoder and aggregation function, generating a new, enhanced feature vector $\hat{\mathbf{h}}$. Finally, the augmented dataset $D_a = \left\{\left(\hat{\mathbf{h}}_{\mathbf{i}}, y_i\right)\right\}$ is used to train the model, improving generalization and robustness through more varied tail label representations.

4 Experiments

4.1 Experiment Setup

Datasets. Experiments were conducted on four well-known benchmark multi-label short text datasets. The statistics for each dataset are summarized in Table 1.

Table 1. Data Statistics. N_{Train}, N_{Test}, W and L refer to the number of training and testing instances, the average number of words per text and the quantity of labels, respectively.

Dataset	N_{Train}	N_{Test}	W	L
Amazon	10,955	4,582	19.63	338
Comment	12,998	3,257	20.14	6
Tweet	7,722	3,258	15.68	11
NTCIR	1,919	639	12.91	8

Evaluation Metrics. In this study, we utilize three widely adopted multi-label evaluation metrics, namely Macro-F1, Micro-F1 and Hamming Loss, to comprehensively assess the performance of our model.

Baselines. We conduct a comprehensive comparison between our proposed KLDAP method and the most representative as well as state-of-the-art (SOTA) methods:

- KSSVPT [12] integrates external knowledge with label prompts to boost short text semantics.
- CoMAL [17] is a framework for multi-label text classification active learning. It devises a sampling strategy to select samples.
- LSFA [18] is a method for long-tailed MLTC. It uses a label-specific encoder for decoupled representation.
- LSAN [19] incorporates label attention and self-attention mechanisms to capture label-specific features.
- LDGN [20] leverages a graph-based structure to capture label-specific semantic interactions among labels for improving classification performance.
- BERT [10] is a highly popular PLM model that can be employed as a text encoder, using fine-tuning to adapt it to MSTC.

- CRFA [21] utilizes multi-stage attention at word and concept levels to enhance short text representations.

Implementation Details. In this study, the performance results for the non-open-sourced KSSVPT method were directly adopted from its original publication while experiments for all other methods were conducted using their official implementations on a standardized dataset. The model was implemented in the PyTorch framework and initialized with the pre-trained BERT-base-uncased model (embedding size 768, subvector size 256), and optimized using Adam. To assess performance in semi-supervised settings, experiments were conducted with 5%, 25%, and 100% labeled data, each repeated five times to report average results.

4.2 Experimental Results and Analysis

Experimental results are evaluated by varying labeling ratios using three metrics (see Fig. 3). Note that the Hamming loss on the Amazon dataset is always below 0.01, making it hard to visualize. After analyzing the results, we make the following observations.

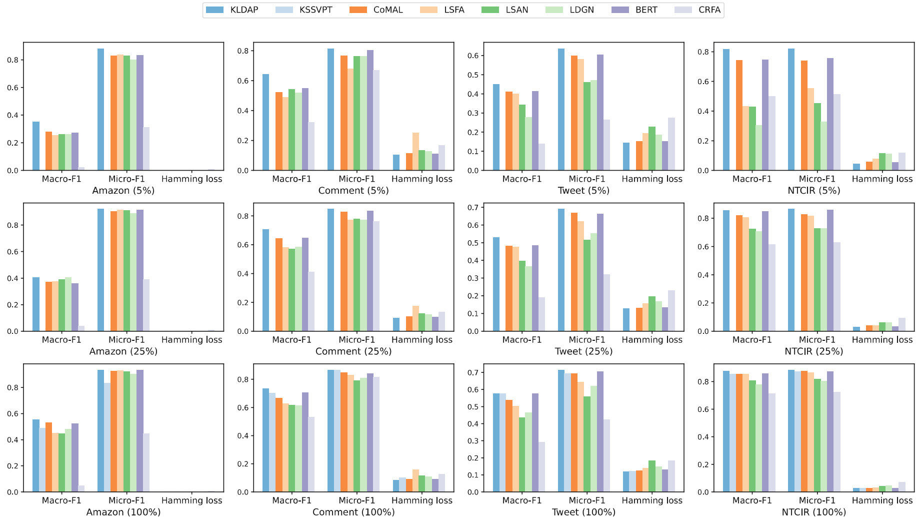

Fig. 3. Experimental results on four datasets.

The influence of the labeling ratio on model performance is empirically substantiated. Specifically, as the proportion of labeled data increases, performance improves consistently. For example, on the Amazon dataset, KLDAP's Macro-F1 score increases from 0.3517 at a 5% labeling ratio to 0.5529 at 100%, highlighting the importance of labeled data in MSTC. In contrast, CRFA underperforms across datasets due to its limited representational capacity and static label embeddings. Notably, at a 5% labeling ratio, KLDAP surpasses the second-best model by 24.89%, 16.97%, 8.72%, and 9.39% in Macro-F1 across four datasets, demonstrating its robustness in low-resource settings.

Furthermore, models like LSFA, LSAN, and LDGN perform well on certain datasets via mechanisms tailored to long-tailed labels, but they fall short in addressing short text sparsity. Similarly, KSSVPT's reliance on fixed semantic mappings reduces its flexibility with complex label structures. Additionally, dataset-specific factors further influence performance. For example, the NTCIR dataset maintains relatively high scores even under low labeling conditions, whereas the Tweet dataset experiences significant performance declines due to higher noise and label sparsity. In summary, by dynamically integrating label dependencies and leveraging external knowledge, KLDAP effectively mitigates challenges such as data sparsity, class imbalance, and complex label interrelations label complexity, thereby establishing itself as a robust approach for MSTC across varying resource scenarios.

4.3 Ablation Study

We evaluated KLDAP through ablation studies by individually removing its three key components: concept extraction (CE), data augmentation (DA), and contrastive learning (CL). Experimental results on Macro-F1 and Micro-F1 scores, are presented in Table 2. The performance degradation observed in each case is analyzed as follows:

Removing the concept extraction module on the Amazon dataset decreased Macro-F1 from 55.29 to 54.86 and Micro-F1 from 93.44 to 93.01, showing its value in addressing data sparsity and enriching semantics. Similarly, excluding data augmentation on the Comment dataset led to a drop in Macro-F1 from 73.51 to 72.06 and Micro-F1 from 86.66 to 85.31, emphasizing its role in boosting training diversity and tail label generalization. Finally, removing contrastive learning on the Tweet dataset reduced Macro-F1 from 57.75 to 56.78 and Micro-F1 from 71.49 to 70.24, demonstrating its critical function in learning label distinctions and improving representation.

In summary, the ablation experiments strongly validate the necessity of the three core modules. Each module makes indispensable contributions to enhancing the model's performance and robustness.

Table 2. Ablation study on four datasets.

Ablation Models	Amazon		Comment		Tweet		NTCIR	
	Ma-F1	Mi-F1	Ma-F1	M-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1
KLDAP	**55.29**	**93.44**	**73.51**	**86.66**	**57.75**	**71.49**	**87.77**	**88.49**
without CE	54.86	93.01	73.24	86.32	57.56	71.08	87.59	88.12
without DA	53.97	92.02	72.06	85.31	56.78	70.24	86.52	87.43
without CL	54.37	92.45	73.04	86.09	57.30	70.72	87.35	87.90

5 Conclusion

In this work, we introduce KLDAP, a novel framework for MSTC tasks that tackles data sparsity, limited labeled data, and class imbalance. By incorporating external knowledge and conceptual enhancement, KLDAP improves the semantic representation of short texts. The pseudo-label optimization strategy effectively leverages unlabeled data, while the VAE module with contrastive learning enhances tail label feature representation. Experiments on four benchmark datasets demonstrate KLDAP's superior performance in handling label sparsity and long-tail distributions. Future work will focus on expanding external knowledge bases, enhancing pseudo-label robustness, and applying the framework to large-scale, real-world tasks.

Acknowledgments. This study was funded by the Natural Science Foundation of China under grants (62376085), Research Funds of Center for Big Data and Population Health of IHM (JKS2023003), National Natural Science Foundation of China (61972425).

Disclosure of Interests. The authors have declared no competing interests.

References

1. Liu, Y., Huang, L., Giunchiglia, F., Feng, X., Guan, R.: Improved graph contrastive learning for short text classification. Proc. AAAI Conf. Artific. Intell. **38**, 18716–18724 (2024)
2. Patel, V., Ramanna, S., Kotecha, K., Walambe, R.: Short text classification with tolerance-based soft computing method. Algorithms **15**, 267 (2022)
3. Tian, G., Wang, J., Wang, R., Zhao, G., He, C.: A multi-label social short text classification method based on contrastive learning and improved ml-KNN. Expert Syst. e13547 (2024)
4. Fiallos, A., Jimenes, K.: Using reddit data for multi-label text classification of twitter users interests. In: 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG), pp. 324–327. IEEE (2019)
5. Deniz, E., Erbay, H., Coşar, M.: Multi-label classification of e-commerce customer reviews via machine learning. Axioms. **11**, 436 (2022)
6. Lu, G., Liu, Y., Wang, J., Wu, H.: CNN-BiLSTM-Attention: a multi-label neural classifier for short texts with a small set of labels. Inf. Process. Manage. **60**, 103320 (2023)
7. Bayer, M., Kaufhold, M.-A., Reuter, C.: A survey on data augmentation for text classification. ACM Comput. Surv. **55**, 1–39 (2022)
8. Falis, M., Dong, H., Birch, A., Alex, B.: CoPHE: A count-preserving hierarchical evaluation metric in large-scale multi-label text classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 907–912, Online; Punta Cana, Dominican Republic (2021)
9. Liu, J., Chang, W.-C., Wu, Y., Yang, Y.: Deep learning for extreme multi-label text classification. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, pp. 115–124 (2017)
10. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186. Minneapolis, Minnesota (2019)

11. Yarullin, R., Serdyukov, P.: BERT for sequence-to-sequence multi-label text classification. In: Analysis of Images, Social Networks and Texts: 9th International Conference, pp. 187–198 (2021)
12. Chen, Z., Li, P., Hu, X.: Knowledge and separating soft verbalizer based prompt-tuning for multi-label short text classification. Appl. Intell. **54**, 8020–8040 (2024)
13. Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C.: Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14471–14481 (2022)
14. Cui, H., Wang, Y., Li, Gangkunand, Welsch, R.E.: Self-training method based on GCN for semi-supervised short text classification. Inform. Sci. **611**, 18–29 (2022)
15. Wang, X., Gao, J., Long, M., Wang, J.: Self-tuning for data-efficient deep learning. In: International Conference on Machine Learning, pp. 10738–10748. PMLR (2021)
16. Yang, W., Zhang, R., Chen, J., Wang, L., Kim, J.: Prototype-guided pseudo labeling for semi-supervised text classification. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 16369–16382. Toronto, Canada (2023)
17. Peng, C., Wang, H., Chen, K., Shou, L., Yao, C., Wu, R., Chen, G.: CoMAL: contrastive active learning for multi-label text classification. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2364–2375 (2024)
18. Xu, P., Xiao, L., Liu, B., Lu, S., Jing, L., Yu, J.: Label-specific feature augmentation for long-tailed multi-label text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10602–10610 (2023)
19. Xiao, L., Huang, X., Chen, B., Jing, L.: Label-specific document representation for multi-label text classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 466–475. Hong Kong, China (2019)
20. Ma, Q., Yuan, C., Zhou, W., Hu, S.: 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, pp. 3855–3864 (2021)
21. Liu, Y., Li, P., Hu, X.: Combining context-relevant features with multi-stage attention network for short text classification. Comput. Speech Lang. **71**, 101268 (2022)