STSPL-SSC: Semi-Supervised Few-Shot Short Text Clustering with Semantic text similarity Optimized Pseudo-Labels

Wenhua Nie[1], **Lin Deng**[2], **Chang-Bo Liu**[1], **Jialing Wei**[3], **Ruitong Han**[4], **Haoran Zheng**[5*]

[1]National Yang Ming Chiao Tung University [2]Macau University of Science and Technology
[3]Shenzhen Technology University [4]University of Bristol [5]University of Auckland

{nie991112.mg11, a544252392}.sc10@nycu.edu.tw
dl16440113120121@gmail.com 202340492002@stumail.sztu.edu.cn
ruitonghan839@gmail.com hzhe810@aucklanduni.ac.nz

Abstract

This study introduces the Semantic Textual Similarity Pseudo-Label Semi-Supervised Clustering (STSPL-SSC) framework. The STSPL-SSC framework is designed to tackle the prevalent issue of scarce labeled data by combining a Semantic Textual Similarity Pseudo-Label Generation process with a Robust Contrastive Learning module. The process begins with employing k-means clustering on embeddings for initial pseudo-Label allocation. Then we use a Semantic Text Similarity-enhanced module to supervise the secondary clustering of pseudo-labels using labeled data to better align with the real clustering centers. Subsequently, an Adaptive Optimal Transport (AOT) approach fine-tunes the pseudo-labels. Finally, a Robust Contrastive Learning module is employed to foster the learning of classification and instance-level distinctions, aiding clusters to better separate. Experiments conducted on multiple real-world datasets demonstrate that with just one label per class, clustering performance can be significantly improved, outperforming state-of-the-art models with an increase of 1-6% in both accuracy and normalized mutual information, approaching the results of fully-labeled classification.

1 Introduction

With Large Language Models (LLM) and Pre-trained Language Models (PLM) advancing rapidly, downstream tasks are increasing, demanding larger datasets, especially in early-stage businesses or specialized domains. These settings often lack labeled data, hindering traditional algorithms. Obtaining task-specific labels is time-consuming and costly, leading researchers to explore unsupervised text clustering. However, such methods require prior knowledge of clustering categories and suffer from uncontrollable clustering centers. Semi-supervised learning under small samples offers a promising solution.

Few-Shot Learning (FSL) (Wang et al., 2020) efficiently categorizes data into meaningful categories with minimal labeled examples. Unlike traditional learning methods that rely on large volumes of labeled data to train models. This is particularly valuable in scenarios where labeled data is scarce or costly to obtain, but unlabeled data is abundant.

Pseudo-labeling generates artificial labels for unlabeled data, aiding training in few-shot learning scenarios (Cascante-Bonilla et al., 2021). This approach leverages the model's own predictions to assign labels to unlabeled instances, effectively using the model's current understanding to augment its training data. In few-shot learning, where labeled examples are minimal, pseudo-label can significantly enhance the learning process by providing a larger, albeit synthetically labeled, dataset. This method allows for iterative refinement of the model's performance, as the pseudo-label data helps bridge the gap between the scarcity of labeled examples and the abundance of unlabeled data. It is particularly valuable in few-shot learning as it circumvents the limitation of having only a few labeled examples, enabling models to learn more complex patterns and improve generalization capabilities.

In this study, we introduce the Semantic text similarity-enhanced Pseudo-Label Enhanced Clustering (STSPL-SSC) framework, a novel semi-supervised learning approach aimed at overcoming the limitations posed by scarce labeled data across various domains. Unlike traditional methods, STSPL-SSC integrates Semantic text similarity-enhanced Pseudo-Label Generation with Robust Contrastive Learning to refine the clustering process effectively. The framework begins by applying k-means clustering on embeddings to generate initial pseudo-labels for each data point. A subse-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2024, pages 12174–12185

quent refinement process, guided by the Semantic text similarity between authentically labeled and pseudo-label data, improves the pseudo-labels' accuracy. This is achieved by employing secondary clustering that not only enhances the clustering effectiveness but also adjusts the pseudo-label to align more closely with the actual clustering centers through Adaptive Optimal Transport (AOT).This is achieved through secondary clustering, which not only improves clustering effectiveness but also adjusts the pseudo-label to align more closely with the actual clustering centers using Adaptive Optimal Transport (AOT). Additionally, STSPL-SSC incorporates a Robust Contrastive Learning module that generates augmentation pairs, facilitating the learning of both categorical and instance-level distinctions. This innovative method significantly bolsters the framework's robustness against imbalanced and noisy datasets, ensuring more reliable clustering outcomes. Through extensive experiments conducted on eight short text clustering datasets, STSPL-SSC demonstrates superior performance, highlighting its effectiveness in semi-supervised learning for short text clustering.

2 Related work

2.1 Semi-Supervised Few-Shot

In the the few-shot scenario, semi-supervised learning is a good solution. Recent research efforts have explored the application of semi-supervised learning to address the few-shot problem: (Hadifar et al., 2019) leveraged an effective Self-Training (ST) method within the realm of semi-supervised learning. Similarly, (Xu et al., 2023) employed LLMs to synthesize data and utilized Self-Training to learn features from the synthesized data. They utilized assignments from a clustering algorithm as supervision to update the weights of the encoder network.

In our research, we opted for real data to ensure minimal errors stemming from external factors. Following the paradigm of Self-Training, we optimize the overall training process using Semantic text similarity.

2.2 Pseudo-Label

Pseudo-label generates predicted labels for unlabeled data, enhancing learning performance with limited annotated data. However, the accuracy of pseudo-labels directly impacts the model's generalization ability, as inaccurate pseudo-labels may lead to performance degradation.

There are several common practices: one method (Wang et al., 2021; Tsai et al., 2022) is based on a self-training strategy, where the basic model is first trained on labeled data and then the model is retrained on unlabeled data and labeled with high-confidence pseudo-labels. Another approach (Sohn et al., 2020) combines the idea of coherence learning, which employs unlabeled data to enhance model robustness under data perturbation. Building on these approaches (Yang et al., 2023)develops previous pseudo-labeling research using prototype learning, which enhances text representations by clustering them using prototypes for low-density separation, and mitigating unbalanced class bias through prototype-guided pseudo-labeling.

In our study, we utilize semantic similarity enhancement and Adaptive Optimal Transport (AOT) to optimize the generation of pseudo-labels, ensuring that the obtained pseudo-labels closely resemble the labeled data.

2.3 Baseline Articles

Our methodology references and improves upon the methods in these two articles. (Zhang et al., 2021) proposed the Supporting Clustering with Contrastive Learning (SCCL) framework, which improves clustering effectiveness by combining self-supervised instance contrastive learning and unsupervised clustering loss. The SCCL model employs pre-trained Sentence Transformer as an encoder and optimizes clustering loss and contrastive loss through end-to-end training. (Zheng et al., 2023) introduced Robust Short Text Clustering (RSTC), which addresses data imbalance and noise issues by introducing Self-Adaptive Optimal Transport (SAOT) and contrastive learning. Our methodology builds upon and enhances the techniques introduced in these two papers. By leveraging semantic similarity enhancement between labeled data and pseudo-labels, we obtain more informative features, thereby improving the effectiveness of subsequent AOT and contrastive learning. Experimental results also validate the feasibility of our approach.

3 Method

3.1 Semantic Textual Similarity Pseudo-Label Semi-Supervised Clustering (STSPL-SSC)

The STSPL-SSC framework introduces an innovative approach to address the challenge of limited

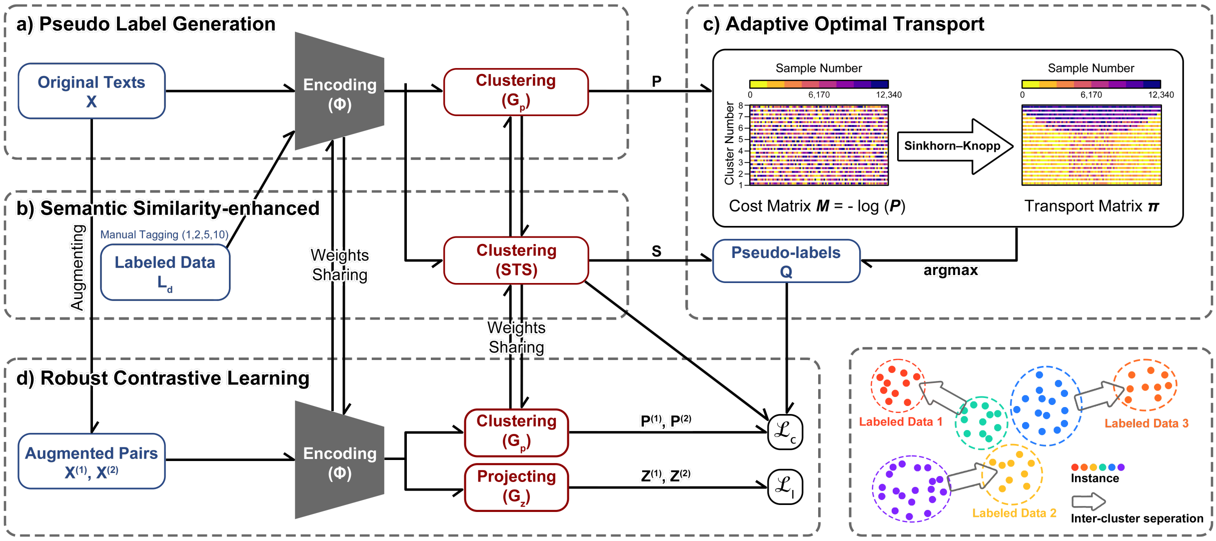


Figure 1: Overall architecture of STSPL-SSC

labeled datasets in various domains, a common obstacle in semi-supervised learning requiring extensive expert tuning (Ren et al., 2018). This framework combines Semantic text similarity enhanced Pseudo-Label Generation with Robust Contrastive Learning, as illustrated in Figure 1. Initially, it employs k-means (MacQueen et al., 1967) clustering on embeddings to assign each data point a pseudo-label P. This is followed by a refinement step where secondary clustering enhances clustering effectiveness, guided by the Semantic text similarity between authentically labeled data L_d, and pseudo-label P. This yields improved pseudo-labels S, while tracking the variances between P and S. To enhance further clustering accuracy, Adaptive Optimal Transport (AOT) is utilized to adjust pseudo-labels Q, closer to the true clustering centers of L_d. Finally, the framework introduces a Robust Contrastive Learning module. This module generates augmentation pairs for each data point, creating augmented batches that facilitate the contrastive learning of categorical and instance-level distinctions. This method improves robustness against imbalanced and noisy data, leading to more stable clustering results.

3.2 Semantic Textual Similarity Pseudo-Label (STSPL)

The Semantic text similarity-enhanced Pseudo-Label Generation module, a cornerstone of the STSPL-SSC framework, aims to address the limitations observed in deep joint clustering methods such as those proposed by (Xie et al., 2016; Hadifar et al., 2019; Zhang et al., 2021; Zheng et al., 2023). Despite their popularity, these methods face challenges primarily due to the lack of supervisory information, which hampers the learning of discriminative representations, and their susceptibility to degenerate solutions, especially in severely imbalanced datasets (Hu et al., 2021; Yang et al., 2017; Ji et al., 2019).

Our module incorporates labeled data L_d during the generation of pseudo-labels P, mirroring a semi-supervised process but eliminating the need for continuous expert optimization of labels. By utilizing L_d as a supervisory signal, we compute the cosine similarity between the embeddings of L_d and the pseudo-label S to gauge their Semantic text similarity. This similarity assessment helps identify the deviation of clustering centers from L_d, thereby enhancing pseudo-label generation and the adaptive optimal transport (AOT) process.

This module unfolds in three primary steps as depicted in Figure 1: Step 1 involves clustering assignment where initial pseudo-labels are assigned. In Step 2, a semi-supervised Semantic text similarity enhancement process leverages the labeled data L_d to refine the pseudo-label S, enhancing their accuracy and relevance. Finally, Step 3 applies AOT to adjust the clustering centers closer to L_d, further refining the pseudo-labels. This approach addresses the challenges of label scarcity and clustering center deviation.

3.2.1 Clustering assignment

The objective of the clustering assignment is to categorize samples with a null label through an initial unsupervised clustering, aiming to derive predictive values for the original texts. To accomplish this, we employ the BGE-M3 model (Xiao et al., 2023) as the encoding network Φ, which is pivotal due to the crucial role of Semantic text similarity enhancement. The effectiveness of utilizing an advanced pre-trained model for word embeddings is confirmed by our experiments. We innovatively combine semantic similarity into the optimization and clustering of pseudo-labels to obtain better clustering results.

The encoding process can be formalized as $E = \Phi(X) \in R^{N \times D_1}$, where X denotes the original text, E the encoded representation, N the batch size, and D_1 the dimensionality of the representation.

Subsequently, a fully connected layer, serving as the clustering network G_p, is utilized to predict the clustering assignment probabilities: $G_p(E) = P \in R^{N \times C}$, where C represents the number of categories. It is essential to highlight that within this module, both the encoding network Φ and the clustering network G_p are kept constant.

3.2.2 Semi-supervised Semantic text similarity enhancement

The aim of semi-supervised Semantic text similarity enhancement is to enhance the clustering assignment outcomes from Step 1. By discerning the extent of deviation from the labeled data, this process attempts to draw cluster centers nearer to the labeled data, hence mitigating the challenges posed by unknown category distributions. Securing more reliable pseudo-labels is a significant concern in such scenarios. Common semi-supervised methods combine supervised learning with unsupervised learning in deep neural networks (Rasmus et al., 2015), or use self-training (ST) (Artetxe et al., 2018; Cai and Lapata, 2019; Gera et al., 2022) approaches typically focus on using student-teacher models to assign pseudo-labels to the unlabelled data, thereby improving accuracy. we compute the cosine similarity between the embeddings of L_d and the pseudo-label P to gauge their Semantic text similarityto get the new pseudo-label S.

The reason for choosing semantic text similarity lies in its similarity to clustering principles, involving computation of vector differences. It is capable of deeply exploring the distances between the pseudo-labels P assigned post-clustering and each labeled data L_d. P will undergo cosine similarity calculation with each L_d to obtain the most similar one, recording the new label as the pseudo-label S. The formula is expressed as follows:

$$S = \text{argmax}_P \left(\frac{P \cdot L_d}{\|P\|_2 ||L_d||_2} \right) \tag{1}$$

3.2.3 Adaptive Optimal Transport (AOT) Method

We refer to the AOT method as outlined in RSTC. The Adaptive Optimal Transport (AOT) method is designed to optimize pseudo-label generation by solving a discrete optimal transport (OT) problem. This process involves several key components and parameters as described below. The AOT optimization problem is formulated as:

$$\min_{\pi, b} \langle \pi, M \rangle + \epsilon_1 H(\pi) + \epsilon_2 (\Psi(b))^T 1 \tag{2}$$

subject to the constraints $\pi 1 = a$, $\pi^T 1 = b$, $\pi \geq 0$, and $b^T 1 = 1$, where the objective function aims to minimize the transportation cost between the probability mass of samples and classes, adjusted by entropy regularization and a penalty function related to class distribution.

After obtaining π, pseudo-labels can be generated via an argmax operation as follows:

$$Q_{ij} = \begin{cases} 1, & \text{if } j = \underset{j'}{\text{argmax}}\, \pi_{ij'}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This operation ensures that each sample is assigned to the class with the highest probability, resulting in a one-hot encoded pseudo-label matrix Q.

Hyperparameters Description:

- ϵ_1 and ϵ_2 are balance hyper-parameters that regulate the impact of entropy regularization and the penalty function, respectively, allowing for a flexible adjustment to accommodate various data characteristics.
- $\Psi(b) = -\log b - \log(1 - b)$ is the penalty function that addresses the distribution of classes by penalizing extreme values of b, thereby encouraging a more uniform distribution of class assignments and preventing clustering degeneracy.
- $H(\pi) = \langle \pi, \log \pi - 1 \rangle$ represents the entropy regularization term, which smoothens the transport plan by discouraging overly sparse solutions, thus facilitating a more robust pseudo-label generation process.

- $a = \frac{1}{N}1$ signifies the uniform distribution of samples, ensuring that each sample contributes equally to the transport process.

- b indicates the initially unknown class distribution.

3.3 Robust Contrastive Learning module

In the Robust Contrastive Learning module, we employ instance augmentation techniques to expand the set of examples and introduce noise to the model, thereby improving its robustness. This is inspired by a body of research that underscores the utility of text augmentation in enhancing model resilience across various settings, as discussed by (Wenzel et al., 2022). Further inspiration comes from (Chen et al., 2020; Zhang et al., 2021; Dong et al., 2022), and the RSTC framework (Zheng et al., 2023), which suggests that post-pseudo-label clustering can exploit instance-level contrastive learning with augmented positive and negative samples to facilitate cluster consolidation and separation.

For implementation, contextual augmenters (Kobayashi, 2018; Ma, 2019) generate two augmented versions of the original text, termed $X^{(1)}$ and $X^{(2)}$. Considering the entire framework utilizes BGE-M3 for embedding analysis, the same method for generating word embeddings is adopted here. This yields augmented representations $E^{(1)}$ and $E^{(2)}$, denoted as $E^{(1)} \in R^{N \times D_1}, E^{(2)} \in R^{N \times D_1}$, where N is the batch size and D_1 is the dimensionality of the embeddings. These are followed by k-means clustering to obtain predicted values $P^{(1)}$ and $P^{(2)}$, expressed as $P^{(1)} \in R^{N \times C}, P^{(2)} \in R^{N \times C}$, where C is the number of clusters. A fully connected layer, serving as the projection network G_z, maps these representations to a new space, facilitating the application of instance-level contrastive loss. The projected representations $Z^{(1)}$ and $Z^{(2)}$ are thus $Z^{(1)} \in R^{N \times D_2}, Z^{(2)} \in R^{N \times D_2}$, with D_2 representing the dimensionality of the new space.

In category-level contrastive learning, we aim for the consistency of cluster predictions between augmentations deemed as positive pairs. Two augmentations from the same original text are treated as a positive pair, with a contrastive task defined on these pairs. The pseudo-label Q serve as the target for the augmented texts, with the L_d acting as the ultimate target. The discrepancy between Q and L_d, represented as α, is calculated as:

$$\alpha = \frac{Q - S}{N} \tag{4}$$

This discrepancy α plays a positive role in the computation of the category-level contrastive loss, which is defined subsequently.

$$\mathcal{L}_C = \alpha \times \frac{1}{N} \left(\|Q - \log P^{(1)}\| + \|Q - \log P^{(2)}\| \right) \tag{5}$$

Instance-level contrastive learning seeks to enhance the consistency between the projection representations of positive augmentation pairs while maximizing the distance between those of negative pairs. For a batch of $2N$ augmented texts, their projection representations are $Z = [Z^{(1)}, Z^{(2)}]^T$. In this batch, for any positive pair (two augmented texts derived from the same original text), the remaining $2(N-1)$ augmented texts are treated as negative samples. The loss function for a positive pair (i, j), where i and j come from the same original text and the rest are considered negatives, is defined as:

$$\ell(i,j) = -\log \frac{\exp(\text{sim}(Z_i, Z_j)/\tau)}{\sum_{k=1}^{2N} 1_{\{k \neq i\}} \exp(\text{sim}(Z_i, Z_k)/\tau)} \tag{6}$$

Within this framework, $\text{sim}(Z_i, Z_j)$ denotes the cosine similarity computed between Z_i and Z_j, and τ is the temperature parameter. The instance-level contrastive loss calculates the loss for all positive pairs within a batch, including both (i, j) and (j, i):

$$\mathcal{L}_I = \frac{1}{2N} \sum_{i=1}^{N} (\ell(i, 2i) + \ell(2i, i)) \tag{7}$$

By integrating pseudo-supervised category-level contrastive learning with instance-level contrastive learning, we are able to derive robust representations that can accurately distinguish between different clusters.

3.4 Overall Framework

The total loss function of the STSPL-SSC model is formulated through a combination of pseudo-supervised class-level contrastive loss and instance-level contrastive loss. Specifically, the overall loss expression of STSPL-SSC is given by:

$$\mathcal{L}_{Total} = \mathcal{L}_C + \lambda_I \cdot \mathcal{L}_I, \tag{8}$$

where λ_I represents a hyperparameter that balances the two types of losses. Adopting this strategy enhances the STSPL-SSC model's ability to handle dataset imbalances and boosts its robustness against data noise. The model consists of two mutually reinforcing modules that form a closed loop, facilitating optimization towards labeled data. As iterations progress, representation learning becomes more robust, and clustering predictions become more accurate, thanks to the more reliable pseudo-labels obtained during the iterative process.

The specific operational procedure is as follows: Initially, we use the k-means clustering algorithm to initialize the embedding, obtaining P, which are then compared with the labeled data L_d to enhance Semantic text similarity, generating pseudo-labels Q. Under the guidance of these pseudo-labels, the model is trained in batches to learn robust representations. Throughout the training process, we dynamically update the Q values using the logarithmic distribution method proposed by (YM. et al., 2020). Finally, by examining the column indices corresponding to the maximum values in each row of the P matrix, we obtain the clustering assignments. Training is terminated when the changes in clustering assignments between two consecutive updates of P are less than a predefined threshold δ, or when the maximum number of iterations is reached. The threshold δ represents the baseline rate of change for the pseudo-labels Q; if this baseline is reached, the optimization improvement is minimal. If the maximum number of iterations is reached without achieving the threshold δ, it indicates that the model may be overfitting, with cluster centers unable to approach the labeled data L_d. Thus, the Adaptive Optimal Transport (AOT) continuously alters the pseudo-labels Q, indicating that no amount of training will result in optimization. This design allows the STSPL-SSC model to self-improve during iterations, optimizing representation and clustering prediction accuracy, thereby achieving higher data processing effectiveness and robustness.

4 Experiments

In this section, we conduct experiments on real-world datasets to emulate the environment encountered in actual work settings. Through these experiments, significant improvements were observed across all datasets, with accuracy (ACC) enhancement rates between 1-7% and Normalized Mutual Information (NMI) enhancement rates also between 1-8%, compared to state-of-the-art short text clustering methods. This illustrates that under the same word embedding model, our Semantic text similarity-enhanced pseudo-label generation module can successfully augment performance, and we have experimentally determined ideal hyperparameters.

4.1 Datasets

Detailed experiments were performed on eight real-world datasets: AgNews, StackOverflow, Biomedical, SearchSnippets, GoogleNews-TS, GoogleNews-T, GoogleNews-S, and Tweet. Among these, AgNews, StackOverflow, and Biomedical are balanced datasets; SearchSnippets is a mildly imbalanced dataset, while GoogleNews, GoogleNews-T, GoogleNews-S, and Tweet are severely imbalanced datasets. Following (Zhang et al., 2021), raw data was used as input to demonstrate our training process's robustness to noise, ensuring a fair comparison. Additional details about the datasets are provided in Appendix A.1.

4.2 Experimental Setup

Our models are implemented in PyTorch 2.0 (Paszke et al., 2019) and trained using the Adam optimizer (Kingma and Ba, 2017).We experimentally selected labeled data number, λ_I, ϵ_1, ϵ_2. More details can be found in Appendix A.2. Following previous studies (Xu et al., 2017; Hadifar et al., 2019; Rakib et al., 2020; Zhang et al., 2021; Zheng et al., 2023), since our method mainly addresses the problem of scarce real data, the number of clusters is set to the actual number of classes, and the evaluation metrics are Accuracy (ACC) and Normalized Mutual Information (NMI). The exact definitions of these metrics are elaborated in Appendix A.3. All experiments are repeated five times and the average results are reported.

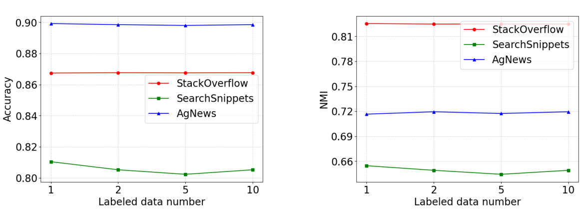
4.3 Baselines

Here is the translation in the style of an academic paper: We compare our proposed method with the following short text clustering techniques and semi-supervised classification methods. SCCL (Zhang et al., 2021) surpasses these approaches by utilizing SBERT (Reimers and Gurevych, 2019) as its backbone and introducing instance-level contrastive learning to support clustering. Moreover, SCCL adopts the deep joint clustering objective proposed by (Xie et al., 2016), obtaining the final clustering assignment via k-means. RSTC

(Zheng et al., 2023) extends SCCL by incorporating a pseudo-label generation module using the SAOT solution and combining it with SCCL's contrastive learning module to enhance robustness against noise. Multi-MCCR (Zhou et al., 2023) is a semi-supervised classification method that consists of multiple models with the same structure and the C-BiKL loss strategy. The C-BiKL loss strategy is proposed to minimize the combination of bidirectional weights. BGE-M3 (Xiao et al., 2023) is a novel self-knowledge distillation approach designed to improve the performance of single retrieval mode. It is the first embedding model that supports all three retrieval methods.

4.4 Clustering Performance

The comparative results on eight datasets are shown in Table 1. From the analysis, we identify several key findings: SCCL achieved improved results by introducing instance-level contrastive learning to mitigate the noise issue but is prone to degenerate solutions and poor application of k-means. RSTC, using SBERT for word embeddings, outperformed previous methods; however, the cluster centers obtained through k-means do not necessarily reflect the labeled data and require iterative refinement, especially for scattered datasets. Multi-MCCR and BGE-M3 simulate the semi-supervised effect by limiting the number of training labels, and their performance is positively correlated with the number of training labels. Table 1 shows the ACC when the number of labels is 100. As they are semi-supervised methods, there is no evaluation of NMI. More experiments are in Appendix A.5. STSPL-SSC surpasses all baselines, demonstrating the effectiveness of enhancing semantic text similarity with labeled data to achieve better clustering performance.



(a) L_d ACC (b) L_d NMI

Figure 3: Impact of L_d number on the model

4.5 In-depth analysis

4.5.1 Ablation Study

To explore the effects of Semantic text similarity and different word embeddings on STSPL-SSC's performance, we compared STSPL-SSC against variants including STSPL-SSC-SS and STSPL-SSC-B. STSPL-SSC-SS utilizes SBERT for word embedding generation, keeping the Semantic text similarity-enhanced pseudo-label generation and Robust Contrastive Learning modules intact. Conversely, STSPL-SSC-B, employing BGE-M3 for word embeddings, excludes the Semantic text similarity enhancement, losing the guidance of labeled data L_d in the pseudo-label Q and loss, maintaining the Robust Contrastive Learning module. It can be observed that both semantic similarity enhancement and the replacement of pre-trained word embedding models have played a significant role. The semantic similarity enhancement module has a notable effect on severely imbalanced models. The reason is that with an increase in the number of dataset categories, semantic similarity enhancement can prevent clustering degradation, thereby improving clustering performance.

4.5.2 Visualization

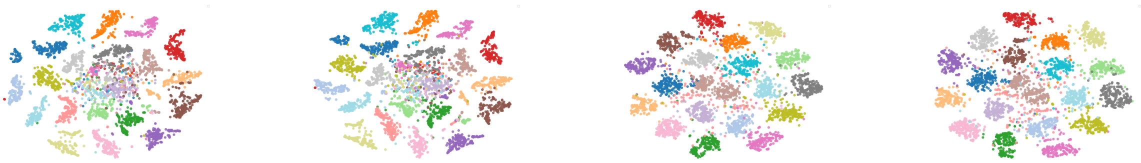
To further demonstrate the effectiveness of the key Semantic text similarity-enhanced module, we employed t-SNE (van der Maaten and Hinton, 2008) for visualizing the representations derived from RSTC, STSPL-SSC-SS, STSPL-SSC-B, and STSPL-SSC. The visualization results on the StackOverflow dataset are depicted in Figures 2(a)-(d). It is evident that: STSPL-SSC achieves the most optimal text representations, characterized by smaller intra-cluster distances and larger inter-cluster distances, with only a minimal number of points misclassified. The underlying reasons for these observations are consistent with the findings analyzed in the ablation study.

4.5.3 Effect of hyper-parameter

We investigate the impact of hyperparameters on model performance, including the number of labeled data, ϵ_1, ϵ_2, and λ_I. Given that the core component is the Semantic Text Similarity-enhanced module, we primarily discuss the influence of the number of labeled data; details on the remaining hyperparameters can be found in Appendix A.4. In datasets where the number of labeled data is sufficient, we experiment with varying the number of labeled data to $\{1, 2, 5, 10\}$, observing negligible

Method	AgNews		SearchSnippets		Stackoverflow		Biomedical	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
SCCL	83.10	61.96	79.90	63.78	70.83	69.21	42.49	39.16
RSTC	85.98	64.32	79.75	69.48	81.97	73.75	43.85	37.99
Multi-MCCR	87.10	-	80.59	-	65.47	-	33.13	-
BGE-M3	87.59	-	80.57	-	75.37	-	47.15	-
STSPL-SSC-SS	85.75	63.53	79.75	**68.68**	83.73	74.25	46.11	38.92
STSPL-SSC-B	89.84	71.39	80.25	64.19	86.53	82.29	47.35	42.28
STSPL-SSC	**89.92**	**71.66**	**81.04**	65.46	**86.74**	**82.54**	**47.43**	**42.49**
Improvement(↑)	2.33	7.34	0.45	-4.02	4.77	8.79	0.28	4.50
Method	GoogleNews-TS		GoogleNews-T		GoogleNews-S		Tweet	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
SCCL	82.51	93.01	69.01	85.10	73.44	87.98	73.10	86.66
RSTC	79.93	92.60	75.50	88.39	76.01	88.27	74.92	85.62
Multi-MCCR	51.42	-	43.33	-	47.32	-	72.34	-
BGE-M3	56.28	-	49.88	-	52.07	-	77.66	-
STSPL-SSC-SS	83.67	93.07	74.94	87.85	78.74	89.39	75.68	85.41
STSPL-SSC-B	**85.15**	**94.36**	78.59	90.77	82.09	**91.54**	70.58	82.02
STSPL-SSC	84.41	94.32	**81.01**	**91.11**	**82.30**	91.18	**79.59**	**88.02**
Improvement(↑)	1.90	1.31	5.51	2.72	6.29	2.91	1.93	1.36

Table 1: Performance comparison across different datasets and methods



(a) RSTC (b) STSPL-SS (c) STSPL-SSC-B (d) STSPL-SSC

Figure 2: t-SNE visualization of the representations on StackOverflow, each color indicates a ground truth category.

impact on balanced datasets with fewer categories, such as AG News and Stack Overflow. However, in the case of SearchSnippets, an increase in labeled data paradoxically led to a decrease in performance, potentially due to the emergence of uncertainty in cluster centroids as the number of labeled data grows, resulting in a deterioration of performance. Based on our experiments, we ultimately opt for number of labeled data $= 1$.

5 Conclusion

This paper presents a robust semi-supervised short text clustering model that includes a Semantic text similarity-enhanced Pseudo Label Generation module and a Robust Contrastive Learning module. Utilizing a semi-supervised approach for generating pseudo labels with labeled data for supervision, our innovation significantly enhances clustering performance by employing few-shot learning to bolster Semantic text similarity, achieving near fully-supervised clustering effectiveness with just one correct label. This greatly increases the usability of unlabeled data for meaningful clustering, reducing costs and providing potential solutions for the lack of training data in downstream tasks of LLM and PLM transfer. Our method demonstrates state-of-the-art performance across eight datasets.

6 Limitations

While the model requires only a minimal number of samples, it still necessitates determining the number of sample categories. Performance degradation can occur when categories have inherently minimal differences, making it challenging for contrastive learning to facilitate cluster separation, potentially leading to data points clustering at the inter-cluster boundaries. Future efforts will focus on overcoming issues of excessive similarity to enhance cluster separation.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Rui Cai and Mirella Lapata. 2019. Semi-supervised semantic role labeling with cross-view training. 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 1018–1027, Hong Kong, China. Association for Computational Linguistics.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6912–6920.

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.

Bo Dong, Yiyi Wang, Hanbo Sun, Yunji Wang, Alireza Hashemi, and Zheng Du. 2022. CML: A contrastive meta learning method to estimate human label confidence scores and reduce data collection cost. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 35–43, Dublin, Ireland. Association for Computational Linguistics.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1119, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amir Hadifar, Lucas Sterckx, Thomas Demeester, and Chris Develder. 2019. A self-training approach for short text clustering. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 194–199, Florence, Italy. Association for Computational Linguistics.

Weibo Hu, Chuan Chen, Fanghua Ye, Zibin Zheng, and Yunfei Du. 2021. Learning deep discriminative representations with pseudo supervision for image clustering. *Information Sciences*, 568:199–215.

Xu Ji, João F Henriques, and Andrea Vedaldi. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, page 91–100, New York, NY, USA. Association for Computing Machinery.

Md Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, and Evangelos Milios. 2020. Enhancement of short text clustering by iterative classification.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. 2018. Meta-learning for semi-supervised few-shot classification.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Austin Cheng-Yun Tsai, Sheng-Ya Lin, and Li-Chen Fu. 2022. Contrast-enhanced semi-supervised text classification with few labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11394–11402.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ximei Wang, Jinghan Gao, Mingsheng Long, and Jianmin Wang. 2021. Self-tuning for data-efficient deep learning. In *International Conference on Machine Learning*, pages 10738–10748. PMLR.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, and Francesco Locatello. 2022. Assaying out-of-distribution generalization in transfer learning. In *Advances in Neural Information Processing Systems*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA. PMLR.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada. Association for Computational Linguistics.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, Jun Zhao, and Bo Xu. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3861–3870. PMLR.

Weiyi Yang, Richong Zhang, Junfan Chen, Lihong Wang, and Jaein Kim. 2023. Prototype-guided pseudo labeling for semi-supervised text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16369–16382.

Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 625–636.

Asano YM., Rupprecht C., and Vedaldi A. 2020. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xiaolin Zheng, Mengling Hu, Weiming Liu, Chaochao Chen, and Xinting Liao. 2023. Robust representation learning with reliable pseudo-labels generation via self-adaptive optimal transport for short text clustering.

Nai Zhou, Nianmin Yao, Qibin Li, Jian Zhao, and Yanan Zhang. 2023. Multi-mccr: Multiple models regularization for semi-supervised text classification with few labels. *Knowledge-Based Systems*, 272:110588.

A Experiments

A.1 Datasets

We conduct extensive experiments on eight popularly used real-world datasets to assess the effectiveness and generality of our approach. The details of each dataset are as follows:

- **AgNews** (Rakib et al., 2020): A subset of AG's news corpus collected by (Zhang et al., 2015), consisting of 8,000 news titles across four topic categories.
- **StackOverflow** (Xu et al., 2017): Comprises 20,000 question titles with 20 different tags, randomly selected from the challenge data published on Kaggle.com.
- **Biomedical** (Xu et al., 2017): Consists of 20,000 paper titles from 20 different topics, selected from the challenge data published on BioASQ's official website.
- **SearchSnippets** (Phan et al., 2008): Contains 12,340 snippets from eight different classes, selected from the results of web search transactions.
- **GoogleNews** (Yin and Wang, 2016): The titles and snippets of 11,109 news articles about 152 events, divided into three datasets: the full dataset is GoogleNewsTS, while GoogleNews-T only contains titles and GoogleNews-S only includes snippets.
- **Tweet** (Yin and Wang, 2016): Consists of 2,472 tweets related to 89 queries, with the original data from the 2011 and 2012 microblog track at the Text Retrieval Conference.

A.2 Experiment Settings

In our experiments, we chose the bge-base-en-v1.5 model (Xiao et al., 2023) from the Sentence Transformer (Reimers and Gurevych, 2019) library for text encoding, with the maximum input length set to 32. The learning rate was set to 5×10^{-6} for optimizing the encoding network, and 5×10^{-4} for optimizing the projection network and clustering network. The dimensions of the text representations and projection representations were set to $D_1 = 768$ and $D_2 = 128$, respectively. The batch size was set to $N = 200$. The temperature parameter was set to $\tau = 1$, and the threshold δ was set to 0.01. For all other baselines, including SCCL (under MIT-0 license) and RSTC, we used the code released by their respective authors.

Table 2: The statistics of the datasets. C means the number of classes, N means the dataset size, A is the average number of words per instance and L/S is the ratio of the size of the largest cluster to that of the smallest cluster

Dataset	**C**	**N**	**A**	**L/S**
AgNews	4	8,000	23	1
StackOverflow	20	20,000	8	1
Biomedical	20	20,000	13	1
SearchSnippets	8	12,340	18	7
GoogleNews-TS	152	11,109	28	143
GoogleNews-T	152	11,108	6	143
GoogleNews-S	152	11,108	22	143
Tweet	89	2,472	9	249

A.3 Evaluation Metrics

We employ two prevalent metrics for evaluating text clustering outcomes: accuracy (ACC) and normalized mutual information (NMI), as adopted by prior research (Xu et al., 2017; Hadifar et al., 2019; Zhang et al., 2021; Zheng et al., 2023).

ACC is given by:

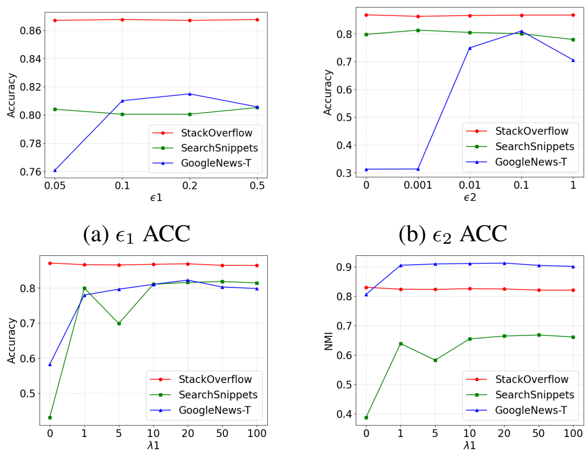
$$ACC = \frac{\sum_{i=1}^{N} 1\{y_i = \text{map}(\hat{y}_i)\}}{N}, \tag{9}$$

where y_i and $\hat{y}_i$ denote the ground truth and the predicted label for the text x_i, respectively.

NMI is given by:

$$NMI(Y, \hat{Y}) = \frac{I(Y, \hat{Y})}{\sqrt{H(Y)H(\hat{Y})}}, \tag{10}$$

where Y and $\hat{Y}$ represent the vectors of ground truth and predicted labels, I denotes the mutual information, and H denotes the entropy.



(a) ϵ_1 ACC (b) ϵ_2 ACC

(c) λ_I ACC (d) λ_I NMI

Figure 4: Impact of hyperparameters on the model

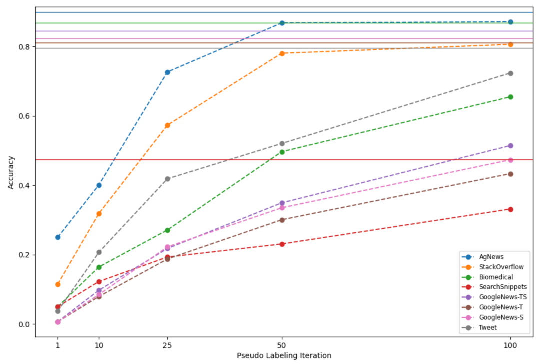


Figure 5: Accuracy Scores for Different Datasets with STSPL-SSC and Multi-MCCR

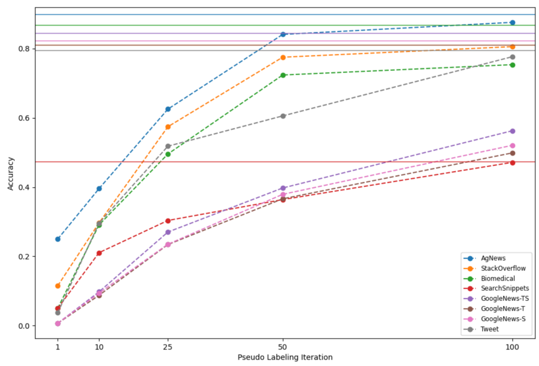


Figure 6: Accuracy Scores for Different Datasets with STSPL-SSC and BGE-M3

A.4 Hyperparametric effect

We investigate the impact of hyperparameters on model performance, including ϵ_1, ϵ_2, and λ_I. We begin by examining the effects of ϵ_1 and ϵ_2, varying them within the sets $\{0.05, 0.1, 0.2, 0.5\}$ and $\{0, 0.001, 0.01, 0.1, 1\}$, respectively. The results are reported in Figures 4(a) and 4(b). Figure 4(a) illustrates that the accuracy is insensitive to ϵ_1. Figure 4(b) highlights the importance of choosing appropriate hyperparameters for datasets with different levels of imbalance, especially for the severely imbalanced GoogleNews-T dataset. Empirically, we select $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.1$ for balanced datasets, $\epsilon_2 = 0.01$ for mildly imbalanced datasets, and $\epsilon_2 = 0.001$ for severely imbalanced datasets. Subsequently, we explore the influence of λ_I by varying it within the set $\{0, 1, 5, 10, 20, 50, 100\}$. The results on three datasets are shown in Figure 4(c) and 4(d). It is observed that performance improves with an increase in λ_I, then remains relatively stable after λ_I reaches 1, and finally decreases when λ_I becomes too large. We conclude that when λ_I is too small, it fails to fully leverage the capabilities of instance-level contrastive learning. Conversely, when λ_I is too large, it suppresses the ability of category-level contrastive learning, thereby diminishing clustering performance. Based on experience, we select $\lambda_I = 10$ for all datasets.

A.5 Comparison with Semi-supervised Methods

In the experiments of Multi-MCCR and BGE-M3 (Figure 5 and Figure 6), referring to the experimental setup of the more advanced semi-supervised method Multi-MCCR, training is performed at the number of labeled data (1, 10, 25, 50, 100), and the training epochs are all 100. The dashed lines represent the training results of the two methods under different label amounts, while the solid lines represent the results of our method. The hyperparameters of Multi-MCCR are set according to the original paper: Hyperp: C = 4, dropout = 0.2, α = 6, Epochs = 8, Maxlength = 128, Batchsize = 12. Under conditions similar to our experiments, our method performs well on severely imbalanced datasets. Due to the large number of categories, if the specific number of categories is uncertain and only partially labeled data is available, it will lead to label imbalance, and the classification situation will be very unsatisfactory. Our method avoids this problem.

A.6 Computational Budget

The training environment we used is the GeForce RTX 4090 GPU, with each dataset taking approximately 15-30 minutes to run.