Graph-augmented and Over-smoothing-resistant Contrastive Clustering for Short Text

Zijian Zheng[1], **Tao Ai**[1], **Yonghe Lu**[1*]

[1]School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China
{zhengzj29, aitao}@mail2.sysu.edu.cn, luyonghe@mail.sysu.edu.cn

Abstract

Short texts present significant challenges for clustering due to semantic sparsity, limited contextual information, and ambiguous category boundaries. While recent studies incorporating contrastive learning and cluster structure optimization have improved performance, their reliance on augmented samples often introduces noise and weakens the capacity of pretrained language models to capture fine-grained semantics. To address these issues, we propose a Graph-augmented and Over-smoothing-resistant Contrastive Clustering framework (GOCC). Specifically, GOCC constructs sentence-level and cluster-level graphs to capture local semantic similarity and global structural patterns, incorporating these signals into sentence representations to enhance representational quality and clustering suitability. Moreover, we introduce a contrastive mechanism based on intermediate layer representations within graph-augmented contrastive learning to alleviate semantic over-smoothing caused by deep networks. Finally, a target-distribution-driven clustering optimization strategy is employed to leverage high-confidence samples in guiding cluster assignments. Experimental results on several benchmark short text datasets demonstrate that GOCC consistently outperforms state-of-the-art methods in terms of clustering accuracy and normalized mutual information.

Code — https://github.com/zjzone/GOCC

Introduction

With the rise of social media platforms and online forums, short texts such as microblogs, user comments, and news summariesb have become increasingly common. Due to their limited length and sparse keyword distribution, these texts often suffer from semantic sparsity, making it difficult to extract latent semantic structures for effective clustering (Lorenzo et al. 2024).

Conventional methods that rely on shallow lexical representations such as TF-IDF and Bag-of-Words (BOW) often struggle to capture semantic meaning because they lack contextual information. In contrast, recent advances in pretrained language models (PLMs), including BERT (Devlin et al. 2019) and Sentence-BERT (SBERT) (Reimers 2019), allow for the extraction of expressive and context-aware sentence embeddings, which provide a more effective foundation for clustering.

However, directly relying on pretrained embeddings without task-specific adaptation often fails to capture subtle semantic variations that are essential for accurate clustering. These embeddings may preserve general linguistic features but lack the discriminative capacity to separate semantically close yet distinct instances. To address this issue, Zhang et al. (2021a) introduce a unified framework that jointly optimizes representation learning and clustering via contrastive learning. By enforcing proximity between semantically similar samples and increasing the distance between unrelated ones in the embedding space, their approach significantly enhances cluster separability. Building on this idea, later studies incorporate additional mechanisms to further refine clustering quality. Neighbor-aware contrastive objectives (Huang et al. 2022) exploit local semantic structures to guide representation learning more effectively. Pseudo-label refinement techniques (Zheng et al. 2023) introduce iterative feedback between clustering assignments and feature learning, leading to more stable and coherent clusters. Moreover, multi-view consistency methods (Zhou et al. 2023) align representations across different augmented views, which reinforces the model's ability to capture invariant and meaningful semantic patterns.

Despite these advancements, several limitations remain. First, most existing approaches conduct contrastive learning at the sequence level, which is prone to introducing redundant noise or semantic drift due to data augmentation, thereby weakening cluster separability. Additionally, these methods primarily rely on pairwise representation optimization and fail to capture more complex relational structures, such as hierarchical semantics and fuzzy cluster boundaries, thus limiting their adaptability and generalizability in scenarios with diverse and ambiguous semantic distributions.

To solve these limitations and leverage both latent cluster structures and the hierarchical semantics of PLMs, we propose GOCC, a novel contrastive clustering framework. GOCC combines multi-view graph representations, hierarchical contrastive learning, and confidence-aware pseudo-label optimization to jointly enhance semantic representations and clustering quality. Specifically, it builds sentence-level and cluster-level semantic graphs to model local and

*Corresponding author

global structures, respectively. These graphs are encoded via relation matrices, and their embeddings are fused through a projection layer for improved structural awareness. To counter semantic drift from data augmentation, GOCC uses intermediate PLM layers as negative samples in a hierarchical contrastive scheme, preserving discriminative features and reducing over-smoothing. A confidence-aware strategy further refines clusters by selecting high-confidence instances to guide pseudo-label updates. In summary, the main contributions of this paper are as follows:

- We propose GOCC, an unsupervised contrastive clustering framework that integrates sentence-level and cluster-level graphs to adaptively refine semantic and structural representations without external knowledge. It effectively addresses diverse semantic distributions and ambiguous category boundaries.
- We introduce a graph-augmented hierarchical contrastive strategy that uses aggregated graph representations as contrastive targets. Intermediate PLM layers enhance representation discrimination and alleviate over-smoothing. A confidence-driven soft clustering module further reinforces cluster structures and promotes clustering-friendly embeddings.
- We conduct extensive experiments on multiple benchmark datasets and demonstrate that the proposed GOCC framework achieves superior clustering performance compared to state-of-the-art methods.

Related Work

Short Text Clustering

Short text clustering is challenging due to limited context, sparsity, and ambiguity. Traditional methods like k-means (Bafna, Pramod, and Vaidya 2016) on term-frequency vectors lack semantic awareness and are often ineffective in sparse and high-dimensional feature spaces. Early neural models (Xu et al. 2017; Hadifar et al. 2019) learn dense embeddings for semantic similarity but don't directly optimize for clustering.

Recent advances introduce contrastive learning guided by clustering tasks (Zhang et al. 2021a), enhancing separability between similar categories. Later methods add pseudo-labeling (Zheng et al. 2023) and self-distillation (Xiao et al. 2024) to refine representations. Yet, most fail to capture fine-grained sentence relations and global cluster semantics, leading to unstable boundary representations and limited clustering quality.

Contrastive Learning

The quality of sentence representations is crucial for short text clustering. Contrastive learning has recently emerged as a powerful self-supervised approach (Wang et al. 2024; Zhang et al. 2021c), encouraging semantically meaningful embeddings by contrasting positive and negative pairs without manual labels. In clustering, most methods use an augmented-view strategy (Zhang et al. 2021a; Yang et al. 2023), treating different augmentations of the same sentence as positives and others as negatives. While effective, such augmentations may distort semantics or introduce noise, harming contrastive signals and clustering performance.

To overcome these limitations, subsequent studies have focused on learning task-specific sentence embeddings through various sampling strategies (Zhang et al. 2022; Deng et al. 2023) and the incorporation of constraints (Lu et al. 2024), aiming to further enhance the expressive power of unsupervised sentence representations. However, the application of augmented view embeddings and different sampling strategies in unsupervised sentence representation learning remains an area that requires further exploration.

Over-smoothing

Over-smoothing in text embedding representations refers to the tendency of token-level embeddings in PLMs to become overly similar, which reduces their semantic discriminability. Prior work (Shi et al. 2022; Li et al. 2025) shows that this issue persists across deeper layers, where node embeddings become increasingly indistinguishable and distort the original semantic structure. Cross-layer over-smoothing further intensifies the problem, as adjacent layers generate highly similar sentence representations (Chen et al. 2023). To address this problem, Chen et al. (2023) incorporate intermediate-layer embeddings as hard negatives in contrastive learning to reduce redundancy and improve representation diversity.

Our work builds on this idea for short text clustering by leveraging intermediate PLM representations as negatives within a graph-augmented contrastive framework. This encourages the model to learn well-separated clusters while preserving fine-grained semantic distinctions.

The Proposed GOCC Method

This section details the GOCC framework, as shown in Fig. 1, which consists of three modules: (i) Graph-augmented representation learning, which enriches sentence embeddings by incorporating local sentence-level semantic and global cluster-level information; (ii) Over-smoothing-resistant contrastive learning, which uses graph-augmented positives and intermediate-layer negatives to boost discriminability; and (iii) Target-distribution-driven clustering, which refines the embedding space via high-confidence samples to improve clustering consistency.

Graph Construction

Although PLM-based sentence embeddings capture rich semantics, they are typically derived from individual sentence contexts and lack explicit modeling of inter-sentence structures or clustering tendencies. To enhance structural awareness, we construct two complementary graphs: a sentence-level graph and a cluster-level graph, which respectively encode local semantic relations and global cluster structure. These structural graphs provide the foundation for subsequent representation learning and clustering optimization. The detailed construction process is as follows.

The sentence-level graph is denoted as $\mathcal{G}_S = \{\mathcal{V}_S, \mathcal{E}_S, \mathbf{X_S}, \mathbf{A_S}\}$, which is constructed from a mini-batch

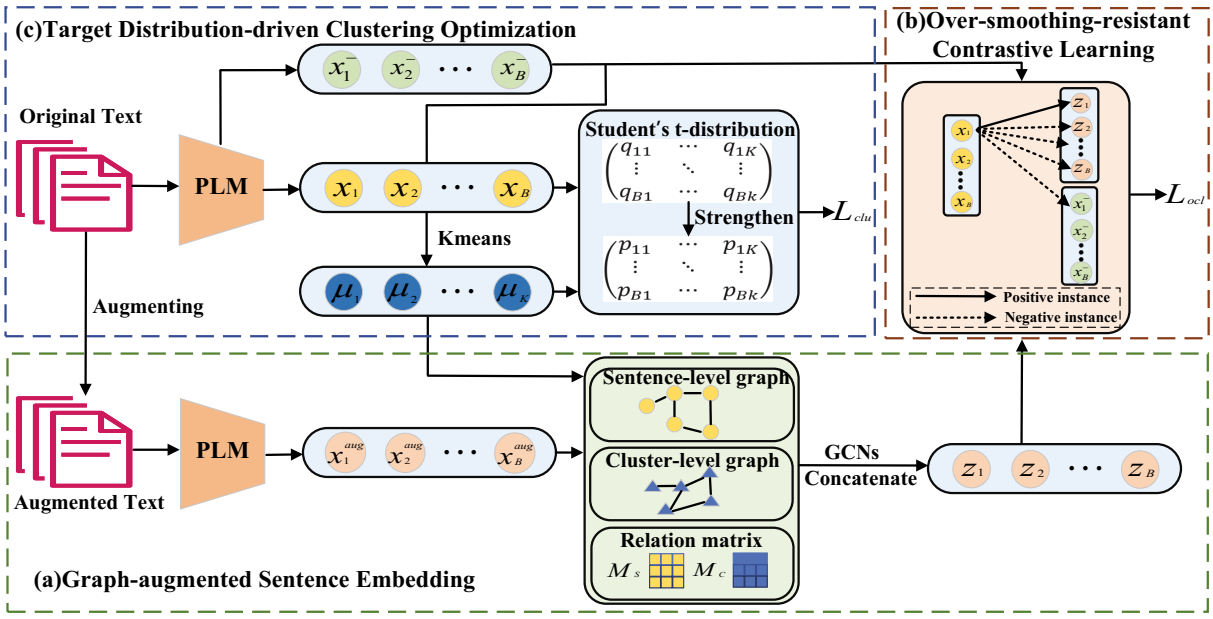


Figure 1: An overall flowchart of GOCC. The framework consists of three stages: (a) constructing two structural graphs $\mathcal{G}_S$ and $\mathcal{G}_C$ for the augmented text along with their relation matrices $\mathbf{M_S}$ and $\mathbf{M_C}$, and obtaining the fused representation $\mathbf{Z}$ via GCNs; (b) performing contrastive learning using the original embeddings $\mathbf{X}$, the graph-augmented representations $\mathbf{Z}$, and intermediate representations $\mathbf{X}^-$ from PLM; (c) computing a cross-entropy loss between the predicted soft assignments and the strengthened target distribution for clustering optimization.

of B text instances. Here, $\mathcal{V}_S = \{{v_S}^1, ..., {v_S}^B\}$ represents the set of sentence nodes, $\mathcal{E}_S$ denotes the set of edges represented by the adjacency matrix $\mathbf{A_S}$, and $\mathbf{X_S} = \{{x_S}^1, ..., {x_S}^B\} \in \mathbb{R}^{B \times d_S}$ contains the sentence embeddings obtained from a pretrained Sentence-BERT model, where d_S denotes the embedding dimension of each sentence. The adjacency matrix $\mathbf{A_S} \in \mathbb{R}^{B \times B}$ encodes pairwise semantic similarity between sentences, where each entry is computed using cosine similarity as: $\mathbf{A_S}_{ij} = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|}$.

The cluster-level graph $\mathcal{G}_C = \{\mathcal{V}_C, \mathcal{E}_C, \mathbf{X_C}, \mathbf{A_C}\}$ is constructed based on the current global cluster centers, introducing high-level clustering structure information into the text representations. Specifically, $\mathcal{V}_C = \{v_C^1, \ldots, v_C^K\}$ represents the set of cluster center nodes, $\mathcal{E}_C$ denotes the edge set represented by the adjacency matrix $\mathbf{A_C}$, and $\mathbf{X_C} = \{{x_C}^1, ..., {x_C}^K\} \in \mathbb{R}^{K \times d_C}$ represents the embeddings of the cluster centers, where the initial representation is obtained by applying the k-means clustering algorithm to the sentence embeddings. To adapt to the continuously optimized representation space, the cluster centers are modeled as learnable parameters and updated during the subsequent target distribution-driven clustering optimization phase, enhancing the model's adaptability to the clustering structure. The adjacency matrix $\mathbf{A_C} \in \mathbb{R}^{K \times K}$ is used to characterize the similarity relationships between the cluster centers, computed using cosine similarity.

Graph-augmented Sentence Embedding

After constructing the sentence-level and cluster-level graphs, we propose a novel graph-augmented sentence embedding to effectively integrate local semantic structures and global clustering relationships in subsequent contrastive learning. Specifically, we use Graph Convolutional Networks (GCNs) to model the feature dependencies and topological structures within the data, which can be defined formally as follows:

$$\mathbf{H}^{(v)} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\mathbf{H}^{(v-1)}\mathbf{W}^{(v)}\right), \tag{1}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ denotes the adjacency matrix with self-loops, ensuring each node incorporates its own features. $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$ is the corresponding degree matrix, and $\mathbf{H}^{(v)}$ denotes the node representations at the v-th GCN layer, with the input $\mathbf{H}^{(0)}$ initialized as $\mathbf{X_S}$ for the sentence-level graph. $\mathbf{W}^{(v)}$ denotes the trainable weight matrix at the v-th layer, and the activation function $\sigma(\cdot)$ is set to ReLU throughout the network.

To more effectively integrate multi-source structural information and enhance the expressiveness of sentence representations, we introduce two relation matrices, $\mathbf{M}_S$ and $\mathbf{M}_C$, which guide structure-aware aggregation over the representations produced by graph convolution. At the sentence-level, we construct $\mathbf{M}_S \in \mathbb{R}^{B \times B}$ with elements defined by the cosine similarity between sentences to capture local semantic relations. At the cluster-level, we introduce $\mathbf{M_C} \in \mathbb{R}^{B \times K}$, representing the soft assignment of each sentence to different cluster centers. Specifically, we use the Student's t-distribution (Xie, Girshick, and Farhadi 2016) to measure the relative distance between sentences and cluster centers, thereby obtaining the corresponding

membership, defined as follows:

$$q_{ik} = \frac{(1 + \|x_S{}^i - x_C{}^k\|_2^2/\gamma)^{-\frac{\gamma+1}{2}}}{\sum_{k'=1}^{K}(1 + \|x_S{}^i - x_C{}^{k'}\|_2^2/\gamma)^{-\frac{\gamma+1}{2}}}, \quad (2)$$

where $x_S{}^i$ is the current embedding of sentence $v_S{}^i$, $x_C{}^k$ represents the embedding of the k-th cluster center, and γ is the degree of freedom of the Student's t-distribution.

Subsequently, we fuse the original sentence representations with the sentence-level and cluster-level structural features propagated through GCNs, resulting in a multi-source sentence representation with both local and global structural awareness. To mitigate redundancy and improve clustering performance, a projection network f_θ is introduced to extract discriminative information from the fused features, producing the final sentence representation for subsequent clustering optimization:

$$\mathbf{Z} = f_\theta([\mathbf{X_S}||\mathbf{M_S} * \mathbf{H_S^{(2)}}||\mathbf{M_C} * \mathbf{H_C^{(2)}}]), \quad (3)$$

where $\mathbf{X_S}$ represents the original pretrained sentence representation, while $\mathbf{H_S^{(2)}}$ and $\mathbf{H_C^{(2)}}$ are the outputs from two-layer GCN encoding over the sentence graph and the cluster graph, respectively. The final fused representation $\mathbf{Z}$ combines sentence semantics with local context and global cluster-level structural information to support robust and clustering-friendly embedding learning.

Over-smoothing-resistant Contrastive Learning

Contrastive learning encourages discriminative representations by maximizing similarity between original and augmented samples while minimizing similarity with others. However, this strategy typically focuses only on sentence-level embeddings from PLMs, ignoring structural relations and clustering characteristics among instances. To overcome this, we adopt $\mathbf{Z}$ as the final representation for augmented instances, allowing the model to capture both local semantic similarity and global clustering structure.

While contrastive learning enhances semantic clarity, it may lead to excessive similarity across hierarchical sentence representations, thereby weakening the PLM's ability to capture fine-grained features. Inspired by SSCL (Chen et al. 2023), we introduce intermediate-layer representations as additional negative samples to promote representational diversity and mitigate semantic over-smoothing.

Specifically, let the original text embeddings of the current batch be $\mathbf{X} = \{x_1, \ldots, x_i, \ldots, x_B\}$, and the embeddings of the augmented view instances be $\mathbf{X^{aug}} = \{x_1^{aug}, \ldots, x_i^{aug}, \ldots, x_B^{aug}\}$. After graph augmentation as shown in Eq. (3), the augmented instance representations are denoted as $\mathbf{Z^{aug}} = \{z_1^{aug}, \ldots, z_i^{aug}, \ldots, z_B^{aug}\}$, and $\mathbf{H}^- = \{x_1^-, \ldots, x_i^-, \ldots, x_B^-\}$ represents the sentence embeddings generated by the intermediate layers of the PLM. The contrastive learning objective for the instance x_i is defined as follows:

$$\ell(x_i) = -\log \frac{e^{\theta(x_i, z_i^{aug})/\tau}}{\sum_{j=1}^{2B} \mathbb{I}_{\{j \neq i\}} e^{\theta(x_i, z_j^{aug})/\tau} + e^{(x_i, x_i^-)/\tau}}, \quad (4)$$

where B is the number of instances in the current batch, $\theta(i, j)$ denotes the cosine similarity between i and j, $\mathbb{I}$ is the indicator function, and τ is the temperature coefficient. The contrastive loss for the entire batch of size B is defined as:

$$\mathcal{L}_{ocl} = \frac{1}{2B} \sum_{i=1}^{B} \ell(x_i) + \ell(z_i^{aug}). \quad (5)$$

It is worth noting that in the denominator of Eq. (4), the intermediate-layer negative samples are taken from the representations of original instances rather than augmented ones. This design prevents semantic drift caused by data augmentation from undermining the model's discriminative ability, thereby mitigating over-smoothing more effectively.

Target Distribution-driven Clustering Optimization

To further improve clustering quality and enhance the tightness between cluster centers and their corresponding instances, we follow the clustering method proposed by Zhang et al (2021a). After obtaining the soft clustering assignment between each instance and each cluster center using the Student's t-distribution from Eq. (2), we construct the target distribution as follows:

$$p_{ik} = \frac{q_{ik}^2 / \sum_{j=1}^{B} q_{jk}}{\sum_{k'} (q_{ik'}^2 / \sum_{j=1}^{B} q_{jk'})}. \quad (6)$$

The target distribution amplifies the impact of high-confidence samples, thereby adaptively improving intra-cluster compactness and inter-cluster separability. To guide the optimization and alignment of cluster centers in the sentence embedding space, we employ a cross-entropy loss to minimize the divergence between the current soft assignments and the target distribution, as defined below:

$$\mathcal{L}_{clu} = -\sum_{i}^{B} \sum_{j}^{K} p_{ij} \log q_{ij}, \quad (7)$$

where B represents the number of samples in the current batch, and K is the total number of clusters.

Model Training and Optimization

Combining Eq. (5) and Eq. (7), the overall loss function of GOCC is defined as follows:

$$\mathcal{L} = \mathcal{L}_{clu} + \lambda \mathcal{L}_{ocl}, \quad (8)$$

where the hyperparameter λ is used to balance weights of the contrastive learning loss and clustering optimization loss.

The training process of GOCC consists of three main stages. First, sentence embeddings from the PLM are refined through joint optimization of sentence-level and cluster-level contrastive learning. Next, in each iteration, updated embeddings and their soft cluster assignments are used to reconstruct the graph structure, enriching the structural signals in augmented samples. Finally, after convergence, the learned representations are clustered via k-means to obtain the final results. The complete training procedure is summarized in Algorithm 1.

Algorithm 1: Training Procedure of GOCC

Input: Unlabeled text corpus $X = \{x_i\}_{i=1}^{n}$; Model parameters θ; Hyperparameters λ, τ, K
Output: Cluster assignments

Initialize the SBERT encoder, clustering centers, and augmented samples X^{aug}.
for each training epoch **do**
  for each mini-batch (X_B, X_B^{aug}) **do**
    Construct the component graphs $\mathcal{G}_S$ and $\mathcal{G}_C$.
    Compute the relation matrices M_S and M_C using cosine similarity and Eq. (2).
    Generate graph-augmented embeddings for the augmented view using Eq. (3).
    Extract intermediate-layer embeddings X^- for contrastive negatives.
    Compute the target distribution p_{ik} based on squared enhancement strategy using Eq. (6).
    Compute contrastive loss $\mathcal{L}_{ocl}$ (Eq. (5)) and clustering loss $\mathcal{L}_{clu}$ (Eq. (7)).
    Update θ by minimizing total loss $\mathcal{L}$ (Eq. (8)).
  end for
end for
Obtain the final sentence representations X^{final} using the trained model.
return $Label \leftarrow k$-means(X^{final}).

Experiments

Experimental Setup

Datasets We evaluate our method on several widely-used benchmark datasets for short text clustering. Dataset statistics are summarized in Table 1, and brief descriptions are provided below:

- **AgNews (AN)** (Zhang, Zhao, and LeCun 2015): A balanced dataset composed of news texts, collected and preprocessed by (Rakib et al. 2020).
- **SearchSnippets (SS)** (Phan, Nguyen, and Horiguchi 2008): A mildly imbalanced dataset consisting of web search snippets across various domains.
- **GoogleNews** (Yin and Wang 2016): Composed of news events, this dataset is divided into full articles, titles, and snippets, denoted as **GoogleNews-TS(G-TS)**, **GoogleNews-T(G-T)**, and **GoogleNews-S(G-S)**, respectively. It is a severely imbalanced dataset.
- **Tweet (TT)** (Yin and Wang 2016): A severely imbalanced dataset of tweets derived from the microblog tracks of a text retrieval conference.

Baseline Methods To verify the effectiveness of the proposed GOCC method, we select a variety of mainstream approaches for comparison, covering different categories of short text clustering techniques. **(I)** Frequency-based methods include **BOW** (Scott and Matwin 1998) and **TF-IDF** (Bafna, Pramod, and Vaidya 2016). **(II)** Representation learning-based methods contain **STCC** (Xu et al. 2017), **Self-train** (Hadifar et al. 2019), **SBERT** (Reimers 2019) and **BGE-M3** (Xiao et al. 2024). **(III)** Contrastive learning-based methods consist of **SCCL** (Zhang et al. 2021a), **PromPos** (Huang et al. 2022) and **CLSESSP** (Shen, Li, and Lin 2024). **(IV)** Semi-supervised and pseudo-label optimization-based methods include **Multi-MCCR** (Zhou et al. 2023) and **RSTC** (Zheng et al. 2023).

Dataset	Classes	Samples	Length(Avg)
AN	4	8000	23
SS	8	12340	18
G-TS	152	11109	28
G-T	152	11109	6
G-S	152	11109	22
TT	89	2472	8

Table 1: Statistics of datasets.

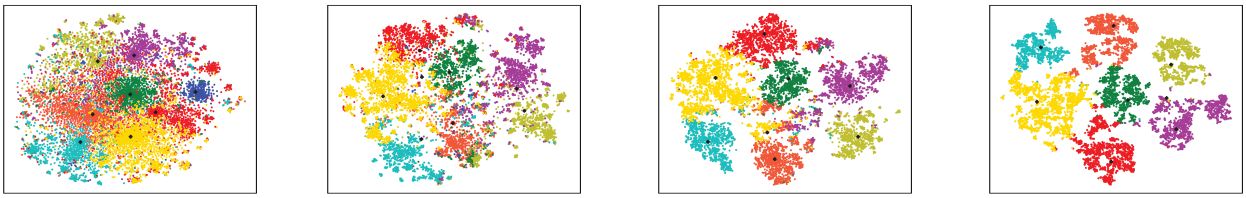
Implementation Details We implement GOCC using the PyTorch framework, with SBERT as the encoder and a two-layer GCN for structure modeling. The embedding dimension is set to 768, and the batch size to 128. We use the Adam optimizer with a learning rate of 1e-5 for the PLM, and 1e-3 for both the contrastive projection head and cluster center optimization. The loss weight λ is set to 7 for all datasets, except for Tweet, where it is set to 1. In the contrastive learning objective, τ is fixed at 0.5. For the soft clustering assignment based on the Student's t-distribution, the degree of freedom γ is set to 1. Additionally, the intermediate layer representations used as negative samples are extracted from the penultimate layer of the PLM.

We apply contextual data augmentation (Kobayashi 2018) to generate one augmented sample for each instance, which has proven effective for short text clustering (Zhang et al. 2021b). Clustering performance is evaluated using Accuracy (ACC) and Normalized Mutual Information (NMI). The specific metric definitions are provided in the Appendix I. Each experiment is run five times with the same settings, and average results are reported for robustness.

Main Results

Table 2 compares the clustering performance of GOCC with multiple baseline models on six short text datasets. Based on this quantitative observation, GOCC achieves the best results in terms of ACC and NMI on all evaluation datasets, demonstrating its superiority in the short text clustering task.

Intriguingly, as a model driven by internal knowledge optimization, GOCC enhances the cluster-friendly sentence distribution produced by PLM through its own embedded information generated in each iteration. This underlines the adaptive ability of the model. Compared to SBERT-based models like SCCL, GOCC's superior performance is attributed to three key designs: (1) multi-granular graph structures enrich structural information absent in plain text embeddings; (2) contrastive learning between graph-augmented and original samples captures both local and global consistency, improving representation discriminability; (3) introducing intermediate-layer PLM representations as negatives mitigates over-smoothing and retains fine-grained semantics.



(a) Raw (b) Iteration 500 (c) Iteration 1000 (d) Final

Figure 2: Clustering visualization results for the SearchSnippets text dataset. Different colors denote distinct clusters and black dots indicate the corresponding cluster centers.

	AN		SS		G-TS		G-T		G-S		TT	
Model	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
BOW	27.60	2.60	24.30	9.30	57.50	81.90	49.80	73.20	49.00	73.50	49.70	73.60
TF-IDF	34.50	11.90	31.50	19.20	68.00	88.90	58.90	79.30	61.90	83.00	57.00	80.70
STCC	83.50	56.90	76.98	62.56	76.90	80.60	70.17	78.97	75.07	86.84	67.92	86.04
Self-Train	63.60	35.50	72.69	56.74	59.40	79.60	58.17	77.39	59.12	78.54	71.64	87.05
SBERT(k-means)	83.44	57.76	73.02	59.77	67.40	90.47	63.98	86.13	65.87	87.64	62.70	86.80
BGE-M3	87.59	63.58	80.57	67.13	56.28	79.34	49.88	79.41	52.07	79.22	77.66	88.35
SCCL	84.62	62.73	75.86	63.67	79.24	92.31	67.32	84.73	77.25	87.37	75.49	89.06
ProPos	84.30	59.30	74.30	55.20	73.90	90.40	65.41	85.32	75.57	87.19	78.42	88.53
CLSESSP	80.45	55.17	69.85	53.29	64.53	85.37	63.60	85.96	64.64	86.83	57.85	81.52
Multi-MCCR	87.10	64.82	80.59	68.46	51.42	78.98	43.33	71.82	47.32	73.41	72.34	87.19
RSTC	84.24	62.45	80.10	69.74	83.27	93.15	72.27	87.39	76.01	88.27	75.20	87.35
GOCC	**87.68**	**66.12**	**85.27**	**70.73**	**87.63**	**93.77**	**74.24**	**87.54**	**78.36**	**88.63**	**81.84**	**90.35**

Table 2: Clustering performance comparison of different methods on six real short text datasets. The best and the second best results are denoted in bold and underline.

We also observe that models combining pseudo-labels or semi-supervised strategies with contrastive learning, such as RSTC, show strong performance. This suggests that external supervisory signals can better constrain category boundaries and alleviate ambiguity. In contrast, traditional term frequency-based methods perform poorly due to their lack of contextual semantics. Deep representation learning models like STCC and Self-Train better capture semantic features but often neglect local structure and global cluster tendencies, leading to overlapping or dispersed clusters in complex distributions. Contrastive learning methods such as SCCL and ProPos improve sample separability by optimizing positive–negative pair relations. However, they mostly operate at the sentence-level and lack global structural modeling, making them sensitive to noisy and intra-cluster variability. In contrast, GOCC implicitly aligns the global cluster structure and local sample relations through iterative embedding optimization, effectively alleviating these problems while maintaining the advantages of contrastive learning.

To further validate GOCC's ability to capture structure and improve clustering, we visualize the sentence embeddings using t-SNE (Van der Maaten and Hinton 2008) on SearchSnippets. As shown in Fig. 2, the clusters become more compact within and more clearly separated between as training progresses, demonstrating GOCC's effectiveness in enhancing local coherence and global structure.

Ablation Study

To validate the effectiveness of each component in GOCC, we design the following ablation variants: "w/o TCO" removes the target-driven clustering optimization; "w/o GAOS" removes both graph-augmented features and the over-smoothing-resistant mechanism, retaining only standard contrastive learning and clustering optimization; "w/o GA" excludes the graph-augmented sentence representation; and "w/o OS" removes the over-smoothing-resistant module. As shown in Table 3, each component contributes to performance gains, confirming the necessity of the proposed designs.

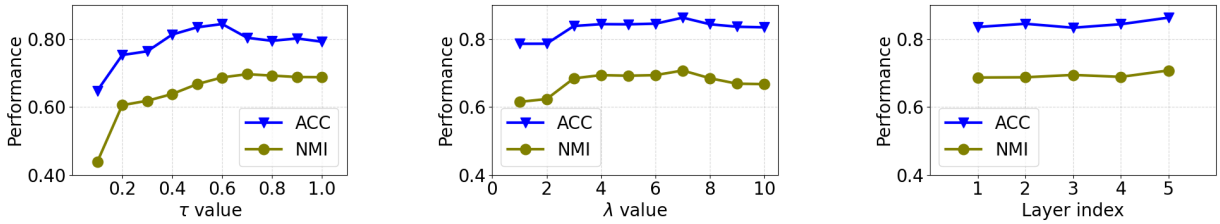
We further ablate the multi-view representations $\mathbf{Z} = f_\theta([\mathbf{X_S}||\mathbf{M_S} * \mathbf{H_S^{(2)}}||\mathbf{M_C} * \mathbf{H_C^{(2)}}])$ in Eq. (3), with results shown in Table 4, where we use $\mathcal{G}_S$ to represent $\mathbf{M_S} * \mathbf{H}^{(2)}$ and $\mathcal{G}_C$ to represent $\mathbf{M_C} * \mathbf{H_C^{(2)}}$. Removing either view leads to consistent performance drops, indicating that both provide complementary structural cues. On the SS dataset, incorporating the sentence-level view improves accuracy by 7.62% over the baseline without graph augmentation. Adding the cluster-level view yields further gains by introducing global structural information that enhances representation discrimination.

Sensitivity Analysis of Hyperparameters

We examine the effects of three key hyperparameters in GOCC: the temperature coefficient τ, contrastive loss

	AN		SS		G-TS		G-T		G-S		TT	
Model	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
w/o TCO	84.76	61.03	78.57	66.87	78.70	92.33	65.57	84.37	66.57	85.33	60.24	83.13
w/o GAOS	84.62	62.73	75.86	63.67	79.24	92.31	67.32	84.73	77.63	87.24	75.49	88.06
w/o GA	86.35	65.48	75.61	62.93	86.32	93.36	71.07	86.94	77.92	87.10	76.46	88.37
w/o OS	86.31	66.04	83.03	65.98	80.95	93.58	67.63	84.36	67.25	83.15	81.51	89.26
GOCC	**87.68**	**66.12**	**85.27**	**70.73**	**87.63**	**93.77**	**74.24**	**87.54**	**78.36**	**88.63**	**81.84**	**90.35**

Table 3: Ablation results of each module in GOCC. Bold values indicate the best results.

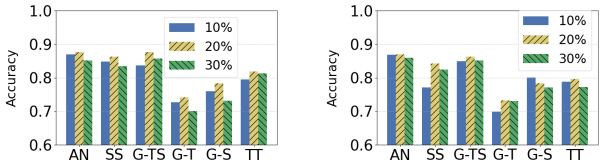


(a) Temperature coefficient (b) Control coefficient (c) Layer index

Figure 3: Effect of varying hyperparameters on GOCC Performance using the SearchSnippets dataset.

$\mathbf{X_S}$	$\mathcal{G}_S$	$\mathcal{G}_C$	AN	SS	G-TS	G-T	G-S	TT
✓	×	×	86.35	75.61	86.32	71.07	77.92	76.46
✓	✓	×	87.06	83.23	87.02	72.37	77.98	79.41
✓	×	✓	86.88	84.16	86.99	71.23	78.04	78.24
✓	✓	✓	**87.68**	**85.27**	**87.63**	**74.24**	**78.36**	**81.84**

Table 4: ACC under different embedding combinations.



(a) BERT (b) RoBERTa

Figure 4: ACC of GOCC when using BERT and RoBERTa as Contextual Augmenters with different substitution rates.

weight λ, and intermediate-layer representations choice.

As shown in Fig. 3 (a), an appropriate τ improves contrastive discriminability, while overly high or low values degrade performance. We set $\tau = 0.5$ for a balance of stability and sharpness. Fig. 3 (b) shows that λ controls the trade-off between local structure modeling and global clustering. Smaller values weaken contrastive signals, while larger ones impair clustering consistency. We use $\lambda = 7$ generally. To assess intermediate-layer negative sampling, we use a 6-layer distilBERT and evaluate representations from layers 1 to 5. Fig. 3 (c) indicates that the fifth layer achieves the best results, supporting prior work suggesting that deeper layers retain more abstract yet discriminative semantic features.

Discussion on Data Augmentation

Given that data augmentation may introduce semantic drift, we investigate the effects of different masked language models and word substitution rates on GOCC using contextual augmentation methods. Although previous work (Zhang et al. 2021a) identified a 20% substitution rate as effective, its applicability within our framework requires validation.

As shown in Fig. 4, we compared the ACC of GOCC using BERT and RoBERTa across different word substitution rates, and observe that BERT with a 20% substitution achieves the best performance. Accordingly, this setting is adopted in our main experiments. Notably, RoBERTa exhibit comparable performance at the same substitution rate, further demonstrating the robustness of this choice.

In addition, we incorporate original and augmented samples into the contrastive learning framework, rather than using only augmented pairs. This design preserves the semantic integrity of the original text while enabling the model to effectively learn relational structures from the constructed graphs. Detailed experiments are provided in Appendix II.

Conclusion

This paper proposes GOCC, a novel short text clustering model that integrates sentence-level and cluster-level graphs to capture both local semantics and global structure. GOCC employs a GCN to encode structural information and generate augmented views, while leveraging intermediate-layer representations from PLM as negative samples to enhance contrastive learning and mitigate over-smoothing. Additionally, it introduces a target distribution-driven clustering objective to refine the representation space. Experiments on multiple benchmark datasets show that GOCC consistently outperforms state-of-the-art methods.

Acknowledgments

This work was supported by the Funding project of The Science and Technology Development Fund of Macao Special Administrative Region (SAR) under the project "Research of AI Key Technologies in Document Mining with Deep Learning and Knowledge Graph" (Grant No. 0018/2024/AMR).

References

Bafna, P.; Pramod, D.; and Vaidya, A. 2016. Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 61–66. IEEE.

Chen, N.; Shou, L.; Pei, J.; Gong, M.; Cao, B.; Chang, J.; Li, J.; and Jiang, D. 2023. Alleviating Over-smoothing for Unsupervised Sentence Representation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3552–3566. Toronto, Canada: Association for Computational Linguistics.

Deng, J.; Wan, F.; Yang, T.; Quan, X.; and Wang, R. 2023. Clustering-Aware Negative Sampling for Unsupervised Sentence Representation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8713–8729. Toronto, Canada: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Hadifar, A.; Sterckx, L.; Demeester, T.; and Develder, C. 2019. A self-training approach for short text clustering. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 194–199.

Huang, Z.; Chen, J.; Zhang, J.; and Shan, H. 2022. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7509–7524.

Kobayashi, S. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 452–457. New Orleans, Louisiana: Association for Computational Linguistics.

Li, J.; Zhang, Q.; Liu, W.; Chan, A. B.; and Fu, Y.-G. 2025. Another Perspective of Over-Smoothing: Alleviating Semantic Over-Smoothing in Deep GNNs. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 6897–6910.

Lorenzo, A. C. M.; Cabot, P.-L. H.; Ghonim, K.; Xu, L.; Choi, H.-S.; Castro, A. F.; and Navigli, R. 2024. Mitigating Data Scarcity in Semantic Parsing across Languages: the Multilingual Semantic Layer and its Dataset. In *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Lu, Y.; Lin, Y.; Yang, M.; Peng, D.; Hu, P.; and Peng, X. 2024. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 14193–14201.

Phan, X.-H.; Nguyen, L.-M.; and Horiguchi, S. 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*, 91–100.

Rakib, M. R. H.; Zeh, N.; Jankowska, M.; and Milios, E. 2020. Enhancement of short text clustering by iterative classification. In *Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems*, 105–117. Springer.

Reimers, N. 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)*, 3982–3992.

Scott, S.; and Matwin, S. 1998. Text classification using WordNet hypernyms. In *Usage of WordNet in natural language processing systems*.

Shen, K.; Li, P.; and Lin, X. 2024. CLSESSP: Contrastive learning of sentence embedding with strong semantic prototypes. *Knowledge-Based Systems*, 299: 112053.

Shi, H.; Gao, J.; Xu, H.; Liang, X.; Li, Z.; Kong, L.; Lee, S. M. S.; and Kwok, J. T. 2022. Revisiting Over-smoothing in BERT from the Perspective of Graph. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, Y.; Yan, X.; Hu, C.; Xu, Q.; Yang, C.; Fu, F.; Zhang, W.; Wang, H.; Du, B.; and Jiang, J. 2024. Generative and contrastive paradigms are complementary for graph self-supervised learning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 3364–3378. IEEE.

Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; and Nie, J.-Y. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 641–649.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, 478–487. PMLR.

Xu, J.; Xu, B.; Wang, P.; Zheng, S.; Tian, G.; and Zhao, J. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88: 22–31.

Yang, X.; Tan, C.; Liu, Y.; Liang, K.; Wang, S.; Zhou, S.; Xia, J.; Li, S. Z.; Liu, X.; and Zhu, E. 2023. Convert: Contrastive graph clustering with reliable augmentation. In *Proceedings of the 31st ACM international conference on multimedia*, 319–327.

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

Zhang, D.; Nan, F.; Wei, X.; Li, S.; Zhu, H.; McKeown, K.; Nallapati, R.; Arnold, A.; and Xiang, B. 2021a. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter ofthe Association for Computational Linguistics: Human Language Technologies*, 5419–5430.

Zhang, D.; Nan, F.; Wei, X.; Li, S.-W.; Zhu, H.; McKeown, K.; Nallapati, R.; Arnold, A. O.; and Xiang, B. 2021b. Supporting Clustering with Contrastive Learning. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5419–5430. Online: Association for Computational Linguistics.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28.

Zhang, Y.; He, R.; Liu, Z.; Bing, L.; and Li, H. 2021c. Bootstrapped unsupervised sentence representation learning. 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)*, 5168–5180.

Zhang, Y.; Zhang, R.; Mensah, S.; Liu, X.; and Mao, Y. 2022. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 11730–11738.

Zheng, X.; Hu, M.; Liu, W.; Chen, C.; and Liao, X. 2023. Robust representation learning with reliable pseudo-labels generation via self-adaptive optimal transport for short text clustering. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, volume 1, 10493–10507.

Zhou, N.; Yao, N.; Li, Q.; Zhao, J.; and Zhang, Y. 2023. Multi-mccr: multiple models regularization for semi-supervised text classification with few labels. *Knowledge-Based Systems*, 272: 110588.