MAST: A Multi-View Alignment Strategy for Optimal Transport-Based Contrastive Clustering of Short Text

Zijian Zheng[1], **Yonghe Lu**[1*]
[1]School of Artificial Intelligence, Sun Yat-sen University, China
{zhengzj29@mail2,luyonghe@mail}.sysu.edu.cn

Abstract

Short text clustering has gained significant prominence due to its ubiquity in real-world applications. Despite the recent success of contrastive clustering, existing paradigms still suffer from two bottlenecks: (1) conventional data augmentation provides limited semantic granularity and may introduce unintended noise; and (2) the absence of global optimization for cluster assignments often precipitates the accumulation of pseudo-label noise, thereby compromising semantic consistency. To bridge these gaps, we propose MAST, a Multi-view Alignment Strategy with Transport-based clustering. MAST constructs complementary structural views to capture multi-granularity semantic features and introduces a multi-view contrastive objective that jointly aligns original, augmented, and structure-enhanced embeddings. To mitigate representation oversmoothing, we incorporate structure-aware negative reweighting and intermediate-layer negative sampling. Furthermore, MAST employs high-confidence guided refinement and an optimal transport-based pseudo-label alignment mechanism to enforce semantic consistency across multiple views. Extensive experiments on several benchmark datasets demonstrate that MAST consistently outperforms state-of-the-art methods, establishing a competitive baseline for short text clustering.

1 Introduction

As a key task in natural language processing and text mining, short text clustering aims to group semantically similar short texts into clusters without manual annotation. This task supports downstream applications such as topic discovery (Meng et al., 2022; Yu and Xiang, 2023) and content retrieval (Zhao et al., 2022; Zheng et al., 2025a). A baseline first represents short texts with frequency-based vectors such as Bag-of-Words (BOW) or TF-IDF (Bafna et al., 2016), and then applies clustering algorithms such as k-means or spectral clustering. However, these representations rely on word occurrence statistics, ignore word order, and fail to capture contextual dependencies. Moreover, as the vocabulary grows, the resulting vectors become high-dimensional and sparse, which makes it difficult to model semantic similarity beyond lexical overlap.

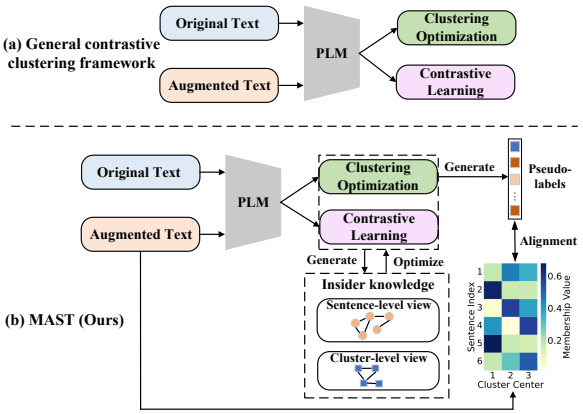


Figure 1: Comparison between existing approaches and MAST. (a) Existing contrastive clustering frameworks rely mainly on augmented-views without structural guidance. (b) MAST incorporates multi-view structural knowledge and aligns optimal transport-based pseudo-labels from the original-view to improve membership consistency in the augmented-view.

With advances in models, researchers have shifted toward distributed representations to mitigate the sparsity and structural limitations of frequency-based features. Word embedding methods such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) learn distributional co-occurrence patterns from large corpora, mapping words into continuous spaces that better capture semantics and improve cluster separability. Sentence encoders based on pre-trained language models (PLMs), such as BERT (Kenton and Toutanova, 2019), further strengthen this direction by encoding fine-grained contextual information

* Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2026, pages 13483–13500
July 2-7, 2026 ©2026 Association for Computational Linguistics

and providing a basis for short text clustering.

Despite pre-trained sentence encoders, short text clustering remains challenging because these models are optimized for language understanding rather than clustering-oriented representations. They do not explicitly enforce inter-cluster separation and often embed ambiguous texts too closely, resulting in entangled embedding spaces (Xu et al., 2023b; Chen et al., 2024). To address this issue, recent work combines contrastive learning with clustering optimization on top of PLMs (Zhang et al., 2021), forming the general framework in Figure 1(a). This paradigm improves separability through contrastive objectives while iteratively refining representations with clustering signals. Subsequent studies extend this framework. Zheng et al. (2023) introduce clustering-aware sampling and adaptive contrastive objectives to reduce false negatives, while Zheng et al. (2025b) incorporate fuzzy neighborhood information to model soft semantic relations. Although these methods enhance representation structure, a key limitation remains: clustering supervision from original texts is not consistently aligned with augmented-views, causing cross-view inconsistency and unstable cluster assignments.

A closer examination of contrastive clustering frameworks reveals implicit supervision that remains underutilized. Original texts provide relatively reliable semantic cues, whereas augmented-views are guided mainly by contrastive objectives, which induce cross-view mismatch and unstable cluster assignments (Xu et al., 2024). We therefore argue that clustering signals should be explicitly propagated across views and informed by both sentence-level and cluster-level structure. Motivated by this observation, we propose MAST, a multi-view alignment framework roughly illustrated in Figure 1(b). MAST integrates structural and semantic cues via multi-view contrastive learning, where sentence-level and cluster-level views complement the explicit augmented-view under a unified objective. It further applies structure-aware negative reweighting to suppress semantically close negatives and enhance inter-cluster separation. To mitigate over-smoothing in graph-enhanced embeddings, MAST incorporates intermediate-layer representations as additional negatives to preserve fine-grained diversity. During clustering, MAST refines original-view assignments using high-confidence instances and aligns augmented-view assignments via optimal transport (OT)-based pseudo-labels, enforcing consistent cluster structure across views.

Our contributions are summarized as follows:

- We propose a multi-view contrastive alignment module that uses the original-view as an anchor and aligns it with the augmented, sentence-level structural, and cluster-level structural views. Without external knowledge, it adaptively refines semantic and structural representations through cross-view alignment.
- We introduce a robust clustering optimization module that refines soft assignments via high-confidence guidance and aligns augmented-view clustering distributions using OT-based pseudo-labels from the original-view, ensuring cross-view consistency.
- Experiments on six benchmark datasets show consistent improvements over strong baselines, supporting the effectiveness of MAST[1].

2 Methodology

2.1 Overall Objective and Training

MAST optimizes a unified objective $\mathcal{L}$ that combines a multi-view contrastive alignment loss $\mathcal{L}_{\text{align}}$ and a clustering loss $\mathcal{L}_{\text{cluster}}$. The clustering loss consists of a high-confidence guided loss $\mathcal{L}_{\text{hgo}}$ and an OT-based pseudo-label alignment loss $\mathcal{L}_{\text{ot-ce}}$. By coupling multi-view contrastive learning with OT-based refinement, this objective promotes multi-granularity semantic structure while preserving consistent cluster assignments across views:

$$\begin{aligned} \mathcal{L} &= \eta\,\mathcal{L}_{\text{align}} + \zeta\,\mathcal{L}_{\text{cluster}}, \\ \mathcal{L}_{\text{cluster}} &= \mathcal{L}_{\text{hgo}} + \mathcal{L}_{\text{ot-ce}}, \end{aligned} \tag{1}$$

where η and ζ control the contributions of the two losses. We set $\eta = 10$ and $\zeta = 1$. For evaluation, we apply k-means to the learned original embeddings to obtain final assignments. Figure 2 provides an overview of the training pipeline.

2.2 Multi-View Contrastive Representation Learning

Instance-level contrastive clustering relies on augmented sentence views, providing only one semantic granularity and limiting structural modeling in short texts, which leads to unstable assignments under semantic sparsity. MAST constructs sentence-level and cluster-level structural views and performs cross-view contrastive alignment to learn

[1] https://github.com/zjzone/MAST

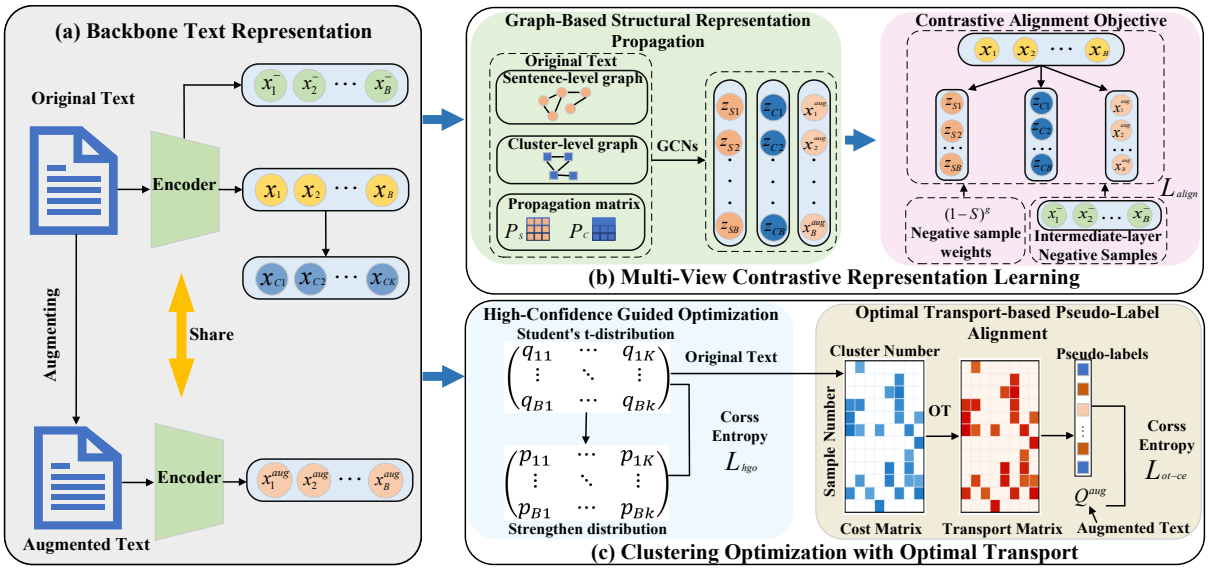


Figure 2: Overview of the proposed MAST framework. (a) Original and augmented texts are encoded into representations $\mathbf{X}^{\text{ori}}$ and $\mathbf{X}^{\text{aug}}$, from which cluster centers $\mathbf{X}_C$ are initialized and intermediate-layer representations $\mathbf{X}^{-}$ are extracted. (b) Sentence-level and cluster-level graphs ($\mathcal{G}_S$, $\mathcal{G}_C$) produce structure-enhanced embeddings $\mathbf{Z}_S$ and $\mathbf{Z}_C$. Using $\mathbf{X}^{\text{ori}}$ as the anchor-view, MAST aligns $\mathbf{X}^{\text{aug}}$, $\mathbf{Z}_S$, and $\mathbf{Z}_C$ via multi-view contrastive learning. (c) Soft assignments are computed via the Student's t-distribution and refined by high-confidence guidance on the original-view; augmented-view memberships are aligned with original-view memberships through OT-based pseudo-label alignment.

multi-granularity representations with improved structural consistency.

Multi-View Structural Modeling. PLM sentence embeddings capture intra-sequence semantics but omit cross-sample relations and inter-cluster structure, which hinders boundary discrimination in short text clustering. MAST therefore constructs two complementary graph views: a sentence-level graph and a cluster-level graph.

The sentence graph $\mathcal{G}_S = (\mathcal{V}_S, \mathbf{X}_S, \mathbf{A}_S)$ is constructed within each mini-batch B to model local semantic relations beyond intra-sequence attention. Here, $\mathcal{V}_S$ is the node set of the B sentences in the batch, $\mathbf{X}_S \in \mathbb{R}^{B\times d_S}$ denotes PLM-encoded sentence embeddings with embedding dimension d_S, and $\mathbf{A}_S \in \mathbb{R}^{B\times B}$ is the adjacency matrix computed by cosine similarity, i.e., $(\mathbf{A}_S)_{ij} = \cos(x_i, x_j)$, where x_i is the i-th row of $\mathbf{X}_S$. Graph propagation reinforces local coherence and yields more discriminative representations.

The cluster graph $\mathcal{G}_C = (\mathcal{V}_C, \mathbf{X}_C, \mathbf{A}_C)$ captures global structure by modeling relations among learnable cluster centers. $\mathcal{V}_C$ is the node set of K cluster centers, and $\mathbf{X}_C \in \mathbb{R}^{K\times d_C}$ denotes center embeddings of dimension d_C, which are initialized via k-means and updated during training. The adjacency matrix $\mathbf{A}_C \in \mathbb{R}^{K\times K}$ is computed by cosine similarity among centers, providing a global relational signal that promotes clearer inter-cluster separation and a coherent embedding geometry.

Graph-Based Structural Representation Propagation. Given the sentence graph $\mathcal{G}_S$ and the cluster graph $\mathcal{G}_C$, MAST applies graph convolutional networks (GCNs) to encode structural dependencies within each graph. The layer-wise update is defined as:

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \quad (2)$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ adds self-loops and $\mathbf{D}_{ii} = \sum_j \hat{\mathbf{A}}_{ij}$ is the degree matrix. $\mathbf{H}^{(l)}$ denotes node representations at layer l, with $\mathbf{H}^{(0)} = \mathbf{X}_S$ for $\mathcal{G}_S$ and $\mathbf{H}^{(0)} = \mathbf{X}_C$ for $\mathcal{G}_C$. $\mathbf{W}^{(l)}$ is a trainable weight matrix and $\sigma(\cdot)$ is ReLU activation function. Through GCN encoding, the model obtains structure-enhanced node representations. We then define two propagation matrices ($\mathbf{P}_S$ and $\mathbf{P}_C$) to encode sentence-level structure and sentence-to-cluster relations, respectively. For the sentence graph, $\mathbf{P}_S \in \mathbb{R}^{B\times B}$ is derived from the normalized cosine similarity adjacency matrix, capturing local semantic relations among sentences. For the sentence-to-cluster relations, $\mathbf{P}_C \in \mathbb{R}^{B\times K}$ is constructed from soft assignment scores between sentences and cluster centers, with $(\mathbf{P}_C)_{ik} = q_{ik}$.

These assignments are computed using a Student's t-distribution (Xie et al., 2016) over Euclidean distances:

$$q_{ik} = \frac{\left(1 + \|x_i - x_{Ck}\|_2^2/\gamma\right)^{-\frac{\gamma+1}{2}}}{\sum_{k'=1}^{K}\left(1 + \|x_i - x_{Ck'}\|_2^2/\gamma\right)^{-\frac{\gamma+1}{2}}}, \quad (3)$$

where x_i denotes the embedding of the i-th sentence, x_{Ck} denotes the embedding of the k-th cluster center, and γ denotes the degrees of the Student's t-distribution. This yields probabilistic soft assignments over clusters, indicating the relative membership of each sentence. Using $\mathbf{P}_S$ and $\mathbf{P}_C$, MAST performs structure-aware aggregation to obtain multi-view enhanced representations, where $\mathbf{P}_S$ aggregates sentence-level features within the mini-batch, while $\mathbf{P}_C$ transfers cluster information to each sentence through soft assignments. The aggregation is defined as:

$$\mathbf{Z}_\pi = \mathbf{P}_\pi \mathbf{H}_\pi^{(2)}, \quad \pi \in \{S, C\}, \quad (4)$$

where $\mathbf{H}_S^{(2)} \in \mathbb{R}^{B\times d_S}$ denotes sentence-node representations and $\mathbf{H}_C^{(2)} \in \mathbb{R}^{K\times d_C}$ denotes cluster-center representations after two GCN layers. Accordingly, $\mathbf{Z}_S \in \mathbb{R}^{B\times d_S}$ and $\mathbf{Z}_C \in \mathbb{R}^{B\times d_C}$ are sentence-level representations aggregated from the sentence graph and the cluster centers, respectively. This aggregation provides each short text with complementary structural signals from sentence-level and cluster-level perspectives, capturing local semantics and global cluster structure.

Contrastive Alignment Objective. Unlike prior contrastive clustering methods that rely only on original and augmented views, MAST derives complementary representations via multi-graph modeling, including the original semantic view, the augmented-view, and structure-enhanced views from sentence-level and cluster-level propagation. These views are projected and normalized into a shared latent space for unified contrastive alignment. Conventional contrastive learning aligns each instance with its augmentation while treating all other samples as negatives, ignoring soft structural relations and often introducing false negatives in short text settings. MAST mitigates this issue through structure-aware negative reweighting based on soft co-clustering relations, which suppresses semantically close negatives and emphasizes truly dissimilar samples. Moreover, inspired by SSCL (Chen et al., 2023), MAST incorporates intermediate-layer representations as additional negatives to counter excessive similarity across representations induced by multi-graph encoding, thereby preserving representational diversity and alleviating semantic over-smoothing.

Specifically, let the original representations in the current mini-batch be $\mathbf{X}^{\text{ori}} = \{x_1, \ldots, x_B\}$ and the explicitly augmented counterparts be $\mathbf{X}^{\text{aug}} = \{x_1^{\text{aug}}, \ldots, x_B^{\text{aug}}\}$. After structural propagation over the sentence-level and cluster-level graphs, the model obtains two structure-enhanced views, $\mathbf{Z}_{\text{S}} = \{z_{\text{S}1}, \ldots, z_{\text{S}B}\}$ and $\mathbf{Z}_{\text{C}} = \{z_{\text{C}1}, \ldots, z_{\text{C}B}\}$. For each instance i, we treat the original representation x_i as the anchor and define the contrasted multi-view representation set as $R_i = \{x_i^{\text{aug}}, z_{\text{S}i}, z_{\text{C}i}\}$. In addition, we extract intermediate-layer sentence embeddings from the PLM as additional negatives, denoted by $\mathbf{X}^- = \{x_1^-, \ldots, x_B^-\}$. To enable structure-aware contrastive alignment in the multi-view embedding space, we compute the soft cluster assignment matrix $\mathbf{Q} \in \mathbb{R}^{B\times K}$ for the original representations $\mathbf{X}^{\text{ori}}$ using Eq. (3). Based on $\mathbf{Q}$, we construct a soft same-cluster relation matrix $\mathbf{S} \in \mathbb{R}^{B\times B}$, where each entry $\mathbf{S}_{ij}$ denotes the probability that samples i and j belong to the same latent cluster:

$$\mathbf{S}_{ij} = \frac{\sum_{k=1}^{K} \mathbf{Q}_{ik}\mathbf{Q}_{jk}}{\sum_{t=1,\, t\neq i}^{B}\sum_{k=1}^{K} \mathbf{Q}_{ik}\mathbf{Q}_{tk}}, \qquad \mathbf{S}_{ii} = 0. \quad (5)$$

This normalization is performed with respect to each anchor instance i, so that S_{ij} reflects the relative same cluster strength of sample j with respect to i among all other samples.

Using the soft structural relation matrix, MAST reweights negative samples and treats all view-specific representations of the same instance as positives. It then defines a multi-view contrastive alignment objective by aggregating positive scores:

$$\mathcal{P}_i = \sum_{r^+ \in R_i} \exp\Big(\frac{\text{sim}(x_i, r^+)}{\tau}\Big), \quad (6)$$

$$\mathcal{N}_i = \underbrace{\sum_{j\neq i}\sum_{u\in R_j} (1 - \mathbf{S}_{ij})^g \exp\Big(\frac{\text{sim}(x_i, u)}{\tau}\Big)}_{\text{Structure-weighted negative samples}} + \underbrace{\sum_{x_j^- \in \mathbf{X}^-} \exp\Big(\frac{\text{sim}(x_i, x_j^-)}{\tau}\Big)}_{\text{Intermediate layer negative samples}}, \quad (7)$$

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\mathcal{P}_i}{\mathcal{P}_i + \mathcal{N}_i}, \quad (8)$$

where τ denotes the temperature parameter controlling the sharpness of the similarity distribution, and g is the negative sample reweighting coefficient. The function $\mathrm{sim}(\cdot,\cdot)$ represents cosine similarity.

It is worth noting that in the multi-view set $R_i = \{x_i^{\text{aug}}, z_{Si}, z_{Ci}\}$, where x_i^{aug} denotes the augmented-view generated by a BERT-based masked language model (MLM) and encoded by the PLM (see detailed analysis in Appendix H). The representations z_{Si} and z_{Ci} are structure-enhanced views obtained by propagating the original representation over the sentence-level and cluster-level graphs, respectively. Intermediate-layer negatives $\mathbf{X}^-$ are collected from an intermediate PLM layer for samples in the mini-batch. This design limits augmentation-induced semantic drift, preserves cross-view semantic consistency, and mitigates over-smoothing during multi-level alignment.

2.3 Clustering Optimization with Optimal Transport

Although contrastive learning improves representations, it does not ensure convergence of cluster centers, as stochastic perturbations during training may distort soft assignment distributions and cause assignment inconsistency. To mitigate this issue, MAST introduces two clustering optimization components: (1) high-confidence guided target updates for stable cluster-center learning, and (2) OT-based pseudo-label alignment to enforce consistency between original and augmented views.

High-Confidence Guided Optimization. After computing soft assignments q_{ik} to cluster centers via Eq. (3), MAST constructs a stabilized target distribution for cluster center optimization. Stochastic variations during training may cause fluctuations in q_{ik}, leading to center drift or oscillation. Following Zhang et al. (2021) and Zheng et al. (2025b), MAST applies squared sharpening to emphasize high-confidence instances while suppressing ambiguous ones, yielding a more reliable target distribution:

$$p_{ik} = \frac{q_{ik}^2 / \sum_{j=1}^{B} q_{jk}}{\sum_{k'=1}^{K} \left(q_{ik'}^2 / \sum_{j=1}^{B} q_{jk'} \right)}. \tag{9}$$

To guide cluster centers toward a stable configuration in the embedding space, we enforce consistency between the current soft assignments and the target distribution using a cross-entropy objective:

$$\mathcal{L}_{\text{hgo}} = -\sum_{i=1}^{B}\sum_{k=1}^{K} p_{ik} \log q_{ik}, \tag{10}$$

where B is the mini-batch size and K denotes the number of clusters.

OT-based Pseudo-Label Alignment. High-confidence guidance stabilizes the original-view assignments, but the augmented-view assignments may still deviate from the underlying cluster structure under contrastive training, introducing noise and assignment drift. To address this issue, MAST formulates cross-view pseudo-label alignment as a discrete OT problem, which performs distribution-level alignment under global constraints and sample-wise transport costs, thereby mitigating assignment imbalance and noise propagation. Using Eq. (3), we first obtain the original-view soft assignment matrix $\mathbf{Q} \in \mathbb{R}^{B\times K}$. Motivated by Zheng et al. (2023), MAST formulates pseudo-label generation as an OT problem and derives an OT-constrained hard pseudo-label matrix $\hat{\mathbf{Y}} \in \{0,1\}^{B\times K}$, where each row indicates the most plausible cluster assignment for an original-view sample. These pseudo-labels are then used to guide the augmented-view soft assignments computed by Eq. (3) through a cross-entropy objective:

$$\mathcal{L}_{\text{ot-ce}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} \hat{y}_{ik} \log q_{ik}^{\text{aug}}, \tag{11}$$

where $\hat{y}_{ik}$ denotes the OT-derived pseudo-label and q_{ik}^{aug} denotes the augmented-view soft assignment probability. This objective enforces the augmented-view to follow the cluster structure induced by the original-view, improving cross-view clustering consistency in the multi-view representation space. Details of the OT formulation are provided in Appendix A.

3 Experiments

3.1 Experimental Setup

3.1.1 Dataset

We evaluate MAST on six widely used short text clustering benchmarks: AgNews (AN), SearchSnippets (SS), Googlenews-TS (G-TS), Googlenews-T (G-T), Googlenews-S (G-S), and Tweet (TT). These datasets cover diverse domains, including news articles, search queries, and social media texts. Detailed statistics are reported in Appendix B.

3.1.2 Baselines

To evaluate the effectiveness of MAST, we compare it with representative short text clustering baselines from four categories:

(I) Traditional frequency-based methods. These methods rely on shallow textual representations derived from word frequency statistics, including BOW and TF-IDF (Bafna et al., 2016).

(II) Deep representation learning methods. These approaches obtain sentence embeddings from PLMs and apply clustering on top of them, including Self-Train (Hadifar et al., 2019), STCC (Xu et al., 2017), Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), and BGE-M3 (Xiao et al., 2024).

(III) Pseudo-label and semi-supervised optimization methods. These methods refine cluster assignments through iterative pseudo-label or label distribution constraints, including RSTC (Zheng et al., 2023) and Multi-MCCR (Zhou et al., 2023).

(IV) Contrastive learning-based clustering methods. These approaches enhance representation separability via contrastive objectives, but typically rely on explicit augmentations or single-view alignment, including SCCL (Zhang et al., 2021), ProPos (Huang et al., 2022), CLSESSP (Shen et al., 2024), and FNSCC (Zheng et al., 2025b).

3.1.3 Implementation Details and Training Process for MAST

We implement MAST in PyTorch by adopting SBERT (Reimers and Gurevych, 2019) as the backbone encoder and a two-layer GCN for structural modeling. The Contextual Augmenter applies 20% word replacement using a BERT-based masked language model. We set the maximum sequence length to 32 and use an embedding dimension of 768. We train MAST for 4000 iterations with batch size $B = 128$ using the Adam optimizer. The learning rate is 1×10^{-5} for SBERT and 1×10^{-3} for the projection head and the learnable cluster centers. We fix $\tau = 0.5$ and set $\eta = 10$, $\zeta = 1$, $\gamma = 1$, and $g = 1$. For the structural views, we set $d_S = d_C = d = 768$, i.e., the input and hidden dimensions in the GCN are the same. For OT, we set $\epsilon_1 = 0.1$. We set $\epsilon_2 = 0.1$ for AgNews, 0.01 for SearchSnippets, and 0.001 for GoogleNews-TS/T/S and Tweet, with 10 solver iterations.

All experiments are repeated five times with identical settings; we report the average performance to ensure robustness. The overall training procedure is summarized in Algorithm 1.

Detailed descriptions of all baseline methods are provided in Appendix C.

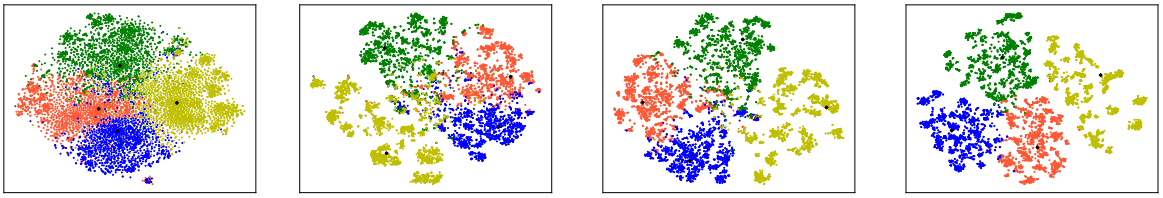
Algorithm 1 Training procedure of MAST.

Input: Unlabeled corpus X, model parameters θ, number of clusters K, and hyperparameters $\eta, \zeta, \gamma, g, \tau$.
Output: Cluster assignments label.

```
Initialize the PLM encoder, cluster centers, and the augmentation module.
for each training epoch do
    for each mini-batch (X_B, X_B^aug) do
        Construct the sentence-level graph G_S and the cluster-level graph G_C.
        Construct P_S and P_C using cosine similarity and Eq. (3).
        Obtain structure-enhanced embeddings Z_S and Z_C via Eq. (4).
        Extract intermediate-layer embeddings X^- from the PLM as additional negatives.
        For each instance i, treat x_i as the anchor and set R_i = {x_i^aug, z_Si, z_Ci}. Compute L_align using Eqs. (5)-(8).
        Construct the target distribution P from Q^ori using Eq. (9) and compute L_hgo via Eq. (10).
        Solve the OT problem on Q^ori to obtain pseudo-labels Ŷ.
        Compute L_ot-ce via Eq. (11).
        Compute the total loss (Eq. (1)) and update θ and the cluster centers.
    end for
end for
Extract the final sentence embeddings X^final from the trained encoder.
return Label ← k-means(X^final).
```

3.2 Main Results

Overall Performance under Data Imbalance and Semantic Sparsity. Table 1 compares MAST with a wide range of state-of-the-art baselines using clustering Accuracy (ACC) and Normalized Mutual Information (NMI), with metric definitions provided in Appendix D. MAST achieves the best results on all six datasets. On the relatively balanced AgNews dataset, it improves ACC by about 0.5 points over the strongest baseline, FNSCC. More importantly, the performance gap increases with class imbalance: on the highly skewed G-T, G-S, and TT datasets, MAST yields an average ACC gain of 4.3 points, demonstrating strong robustness to cluster-frequency bias. From the perspective of



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

Figure 3: Clustering visualization on the AgNews dataset. Different colors denote different clusters, and black dots indicate 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
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
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
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
RSTC	84.24	62.45	80.10	69.74	83.27	93.15	75.50	88.39	76.01	88.27	75.20	87.35
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
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
FNSCC	87.85	66.70	82.59	67.65	88.21	94.31	72.72	87.76	80.49	89.37	83.62	90.38
MAST	**88.34**	**68.39**	**84.98**	**69.94**	**88.34**	**94.57**	**80.14**	**89.65**	**82.60**	**89.81**	**90.41**	**93.87**

Table 1: Clustering performance (%) on six short text benchmark datasets. Best and second-best results are highlighted in **Bold** and underlined, respectively. Improvements over the strongest baseline are statistically significant under a paired t-test ($p < 0.005$).

text length, AN, G-TS, and G-S consist of short but semantically complete news-style texts, whereas G-T and TT contain extremely short texts with highly sparse semantics, with an average length of 6 to 8 words. MAST exhibits larger improvements in these more challenging settings, suggesting that multi-view structural modeling and OT-based pseudo-label alignment effectively enhance clustering consistency under weak semantic signals and fuzzy boundaries.

Comparison across Method Categories. We further analyze the performance across different categories of baselines. Frequency-based methods perform poorly due to their inability to capture contextual semantics. Deep representation methods such as STCC and Self-Train encode sentence-level semantics but fail to model local structure and global cluster organization, often resulting in overlapping clusters. Contrastive clustering methods, including SCCL and ProPos, improve separability, yet their single-view objectives and lack of global assignment constraints make them sensitive to noise and semantic sparsity. Methods with pseudo-label refinement or semi-supervision, such as RSTC, perform better, underscoring the importance of assignment guidance. In contrast, MAST integrates multi-view structural modeling, structure-aware contrastive alignment, and OT-based pseudo-label refinement, achieving optimal performance across all datasets.

Visualization of Clustering Dynamics. To further examine the optimization dynamics of MAST, Figure 3 presents t-SNE visualizations (Fujiwara et al., 2021) of sentence embeddings at different training stages. As training progresses, inter-cluster separation increases while intra-cluster representations become more compact, indicating that MAST gradually forms clearer and more stable clustering structures.

Additional Experimental Results. Further analyses and discussions are provided in Appendices E-I, including iterative visualization on SearchSnippets (Appendix E), comparisons of graph encoder variants for structural views (Appendix F), a case study with error analysis (Appendix G), an evaluation of data augmentation settings (Appendix H),

	AN		SS		G-TS		G-T		G-S		TT	
Model	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
w/o CAO	82.55	59.22	69.44	49.64	78.06	90.55	65.59	86.30	68.35	85.09	86.65	91.64
w/o NSRC	87.10	66.43	84.62	69.12	87.05	93.79	77.84	89.11	81.01	89.08	88.23	93.45
w/o INS	87.68	67.89	84.68	68.69	87.04	94.07	78.83	89.25	81.13	89.34	89.44	93.33
w/o HGO	86.44	66.57	79.10	67.38	74.36	91.44	70.65	87.81	68.54	85.79	70.23	87.47
w/o OTPA	87.90	68.05	83.65	68.13	87.13	93.88	77.41	89.08	82.29	89.45	88.67	93.06
MAST	**88.34**	**68.39**	**84.98**	**69.94**	**88.34**	**94.57**	**80.14**	**89.65**	**82.60**	**89.81**	**90.41**	**93.87**

Table 2: Ablation results (%) of individual components in MAST. **Bold** values indicate the best results.

and a discussion on training time and computational complexity (Appendix I).

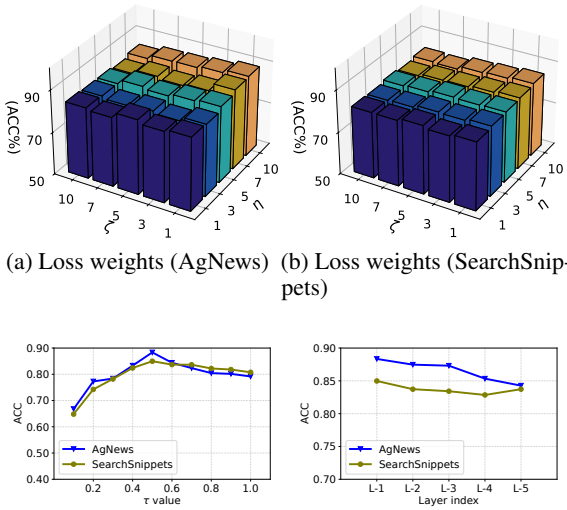
3.3 Ablation Study

To examine the contribution of each component in MAST, we perform ablation studies on six short text datasets and evaluate five model variants: (1) **w/o CAO**, which removes the multi-view contrastive alignment objective; (2) **w/o NSRC**, which removes the structure-aware negative reweighting component in CAO; (3) **w/o INS**, which removes intermediate-layer negative samples in CAO; (4) **w/o HGO**, which removes the high-confidence guided optimization module; (5) **w/o OTPA**, which removes the OT-based pseudo-label alignment component. Table 2 compares these variants with the full MAST model. Removing any component consistently degrades performance across datasets, indicating that each module contributes to the overall effectiveness of the framework. For **w/o CAO**, we remove the entire alignment objective, including NSRC and INS.

In addition to these core component ablations, we further analyze the role of graph-structured views in the multi-view contrastive alignment objective by separately removing the sentence-level view $\mathbf{Z}_S$ and the cluster-level view $\mathbf{Z}_C$. As shown in Table 3, both views provide consistent performance gains. This result suggests that local semantic relations and global cluster-level structure captured by the graphs complement the information missing from the augmented-view, which is particularly beneficial under the semantic sparsity of short texts.

$\mathbf{X}^{aug}$	$\mathbf{Z}_S$	$\mathbf{Z}_C$	AN	SS	G-TS	G-T	G-S	TT
✓	×	×	87.56	83.42	86.87	78.23	81.54	88.23
✓	✓	×	87.93	84.14	87.32	78.89	81.91	89.14
✓	×	✓	88.05	84.32	87.17	79.54	81.77	89.04
✓	✓	✓	**88.34**	**84.98**	**88.34**	**80.14**	**82.60**	**90.41**

Table 3: ACC (%) under different embedding combinations for multi-view contrastive learning. **Bold** indicates a better performance.



(a) Loss weights (AgNews) (b) Loss weights (SearchSnippets)

(c) Temperature τ (d) Intermediate-layer depth

Figure 4: Hyperparameter sensitivity analysis of MAST on the AgNews and SearchSnippets datasets.

3.4 Sensitivity Analysis of Hyperparameters

To evaluate the robustness of MAST with respect to key hyperparameters, we analyze the effects of the loss weight ratio between the contrastive alignment and clustering objectives (η and ζ), the temperature parameter τ, and the depth of intermediate-layer negative samples. Experiments are conducted on the AgNews and SearchSnippets datasets.

Figures 4(a) and 4(b) report performance under different loss weight settings. The combination $\eta = 10$ and $\zeta = 1$ achieves the strongest performance across both datasets. Figure 4(c) shows the impact of the temperature parameter τ, where $\tau = 0.5$ yields the highest performance. Extremely large or small values reduce the contrast between positive and negative pairs, leading to performance degradation. Figure 4(d) examines the effect of intermediate-layer negative selection and shows that using representations from the penultimate PLM layer (denoted as $L-1$) yields the most effective results. This finding in MAST is consistent with prior studies (Chen et al., 2023), which report

that deeper layers encode more abstract yet discriminative semantic information and are therefore more suitable for contrastive clustering.

4 Related Work

Contrastive Learning. Contrastive learning has become a core paradigm for self-supervised sentence representation learning, as it pulls semantically related samples closer while pushing unrelated ones apart (Xu et al., 2023a; Yang et al., 2025). Representative methods such as SimCLR (Chen et al., 2020) and SimCSE (Gao et al., 2021) construct positive pairs through augmented views or stochastic perturbations and learn strong label-free embeddings. Contrastive objectives are also widely used in multimodal alignment (Li et al., 2023).

In short text clustering, contrastive learning is commonly adopted to improve embedding separability. However, most existing methods rely on single-view augmentation and treat all non-positive samples as negatives (Zhang et al., 2021; Yang et al., 2023). This design tends to introduce false negatives and exacerbate cross-view mismatch under semantic sparsity. To address these issues, recent studies explore more informed sampling strategies (Zhang et al., 2022) and structural constraints (Deng et al., 2023; Lu et al., 2024) to achieve better task-specific alignment.

Short Text Clustering. Short text clustering remains challenging due to semantic sparsity, limited context, and ambiguous boundaries. Existing methods address these issues only partially and lack a unified treatment of semantic uncertainty. Prior work ranges from lexical approaches based on surface statistics (Bafna et al., 2016) to deep representation models such as Self-Train (Hadifar et al., 2019), STCC (Xu et al., 2017), SBERT (Reimers and Gurevych, 2019), and general purpose embeddings like BGE-M3 (Xiao et al., 2024).

Despite improved semantic encoding, PLMs still fail to capture local structure and global cluster organization. Another line of research combines representation learning with iterative assignment refinement. Pseudo-label methods such as RSTC (Zheng et al., 2023) and Multi-MCCR (Zhou et al., 2023) improve robustness but remain sensitive to noisy assignments. Contrastive clustering frameworks, including SCCL (Zhang et al., 2021), ProPos (Huang et al., 2022), and CLSESSP (Shen et al., 2024), enhance separability but rely on single-view augmentations and uniform negative sampling, which introduce false negatives and cross-view inconsistency. FNSCC (Zheng et al., 2025b) mitigates false negatives via fuzzy neighborhoods, yet it still depends on single-view contrastive learning and lacks fine-grained structural semantics. These limitations motivate MAST, which leverages multi-view structural signals and cross-view consistency learning to improve representation quality and assignment stability.

5 Conclusion

In this paper, we propose MAST, a multi-view alignment framework for short text clustering. MAST addresses two fundamental challenges in this task: the limited semantic granularity of single-view contrastive learning and the instability of cluster assignments caused by noisy pseudo-labels. It formulates short text clustering as a joint optimization problem that couples representation learning with assignment refinement. By integrating explicit augmentation views with sentence-level and cluster-level structural views under a unified contrastive objective, MAST learns more informative and consistent representations. In addition, high-confidence refinement and optimal transport-based pseudo-label alignment further stabilize cluster assignments and improve cross-view consistency. Experiments on six benchmark datasets show that MAST consistently outperforms strong baselines and remains robust under severe semantic sparsity and distribution imbalance.

Limitations

Although MAST demonstrates strong performance across multiple benchmarks, several aspects still offer room for refinement. The integration of multi-view structural modeling introduces modest computational overhead compared with purely instance-level contrastive approaches, while the overall framework remains efficient in practice. MAST also relies on fixed hyperparameters to balance contrastive and clustering objectives, and developing adaptive or data-driven weighting strategies is a promising direction for future work. In addition, the current structural views are constructed using cosine similarity and soft assignments, which capture only coarse relational patterns. Exploring more expressive or learnable graph structures, such as dynamic neighborhood selection or attention-based graph modeling, may further enhance the ability to represent fine-grained semantic dependencies.

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

Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya. 2016. Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE.

Nuo Chen, Linjun Shou, Jian Pei, Ming Gong, Bowen Cao, Jianhui Chang, Jia Li, and Daxin Jiang. 2023. Alleviating over-smoothing for unsupervised sentence representation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3552–3566, Toronto, Canada. Association for Computational Linguistics.

Sihao Chen, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. 2024. Sub-sentence encoder: Contrastive learning of propositional semantic representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1596–1609, Mexico City, Mexico. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Jinghao Deng, Fanqi Wan, Tao Yang, Xiaojun Quan, and Rui Wang. 2023. Clustering-aware negative sampling for unsupervised sentence representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8713–8729, Toronto, Canada. Association for Computational Linguistics.

Yasuhiro Fujiwara, Yasutoshi Ida, Sekitoshi Kanai, Atsutoshi Kumagai, and Naonori Ueda. 2021. Fast similarity computation for t-SNE. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 1691–1702. IEEE.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. 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.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.

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

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Minneapolis, Minnesota.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore. Association for Computational Linguistics.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. 2024. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 14193–14201.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM web conference 2022*, pages 3143–3152.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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*, pages 91–100.

Md Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, and Evangelos Milios. 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*, pages 105–117. Springer.

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.

Kaicheng Shen, Ping Li, and Xiao Lin. 2024. Clsessp: Contrastive learning of sentence embedding with strong semantic prototypes. *Knowledge-Based Systems*, 299:112053.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 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*, pages 641–649.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487. PMLR.

Binbin Xu, Jun Yin, and Nan Zhang. 2024. Graph based consistency learning for contrastive multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8633–8641.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023a. SimCSE++: Improving contrastive learning for sentence embeddings from two perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12028–12040, Singapore. Association for Computational Linguistics.

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

Lingling Xu, Haoran Xie, Zongxi Li, Fu Lee Wang, Weiming Wang, and Qing Li. 2023b. Contrastive learning models for sentence representations. *ACM Transactions on Intelligent Systems and Technology*, 14(4):1–34.

Xihong Yang, Cheng Tan, Yue Liu, Ke Liang, Siwei Wang, Sihang Zhou, Jun Xia, Stan Z Li, Xinwang Liu, and En Zhu. 2023. Convert: Contrastive graph clustering with reliable augmentation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 319–327.

Yang Yang, Wei Shen, Junfeng Shu, Yinan Liu, Edward Curry, and Guoliang Li. 2025. Cmvc+: a multi-view clustering framework for open knowledge base canonicalization via contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2296–2310.

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

Dejian Yu and Bo Xiang. 2023. Discovering topics and trends in the field of artificial intelligence: Using lda topic modeling. *Expert systems with applications*, 225:120114.

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. *Advances in Neural Information Processing Systems*, 28.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11730–11738.

Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 970–981.

Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2025a. PPTAgent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14413–14429, Suzhou, China. Association for Computational Linguistics.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10493–10507, Toronto, Canada. Association for Computational Linguistics.

Zijian Zheng, Yonghe Lu, and Jian Yin. 2025b. FNSCC: Fuzzy neighborhood-aware self-supervised contrastive clustering for short text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2831–2846, Suzhou, China. Association for Computational Linguistics.

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 Optimal Transport

Inspired by RSTC (Zheng et al., 2023), we formulate an OT problem over the soft cluster assignment matrix of the original view, $\mathbf{Q}^{\text{ori}} \in \mathbb{R}^{B\times K}$, to infer cross-view consistent pseudo-labels $\hat{\mathbf{Y}} \in \{0,1\}^{B\times K}$. We define the cost matrix as $\mathbf{C} = -\log \mathbf{Q}^{\text{ori}}$, and introduce a transport matrix $\boldsymbol{\pi} \in [0,1]^{B\times K}$ to model the joint distribution between samples and clusters. The OT objective is formulated as:

$$\begin{aligned} &\min_{\boldsymbol{\pi},\boldsymbol{b}} \langle \boldsymbol{\pi}, \mathbf{C}\rangle - \epsilon_1 H(\boldsymbol{\pi}) + \epsilon_2 \Omega(\boldsymbol{b}), \\ &\text{s.t.} \quad \boldsymbol{\pi}\mathbf{1} = \boldsymbol{a}, \ \ \boldsymbol{\pi}^T\mathbf{1} = \boldsymbol{b}, \ \ \boldsymbol{b}^T\mathbf{1} = 1, \ \ \boldsymbol{\pi} \geq 0. \end{aligned} \tag{12}$$

where $\boldsymbol{a} = \frac{1}{B}\mathbf{1}$ denotes a uniform marginal distribution over samples, and $\boldsymbol{b}$ is a learnable marginal distribution over clusters. To promote stable transport and balanced cluster usage, we introduce an entropy regularization term $H(\boldsymbol{\pi}) = -\sum_{i=1}^{B}\sum_{j=1}^{K}\pi_{ij}\big(\log \pi_{ij} - 1\big)$ and a cluster-marginal regularizer $\Omega(\boldsymbol{b}) = -\sum_{j=1}^{K}\big(\log b_j + \log(1-b_j)\big)$ with ϵ_1 and ϵ_2 controlling their respective strengths.

It is worth noting that π_{ij} represents the amount of probability mass transported from sample i to cluster j. Since cluster imbalance is common in short text scenarios, we model $\boldsymbol{b}$ as adaptive and use $\Omega(\boldsymbol{b})$ to prevent it from collapsing into a single dominant cluster. Using the method of Lagrange multipliers, the constrained OT problem above can be rewritten equivalently as the following optimization form:

$$\begin{aligned} \min_{\boldsymbol{\pi},\boldsymbol{b}} \ L = &\langle \boldsymbol{\pi}, \mathbf{C}\rangle - \epsilon_1 H(\boldsymbol{\pi}) + \epsilon_2 \Omega(\boldsymbol{b}) - \\ &\boldsymbol{f}^T(\boldsymbol{\pi}\mathbf{1} - \boldsymbol{a}) - \boldsymbol{g}^T(\boldsymbol{\pi}^T\mathbf{1} - \boldsymbol{b}) - h(\boldsymbol{b}^T\mathbf{1} - 1), \end{aligned} \tag{13}$$

where $\boldsymbol{f}$, $\boldsymbol{g}$, and h denote the Lagrange multipliers that incorporate the OT constraints into the unconstrained optimization formulation. Taking the derivative of Eq. (13) with respect to π_{ij} and setting $\frac{\partial L}{\partial \pi_{ij}} = 0$ yields:

$$\pi_{ij} = \exp\left(\frac{f_i + g_j - C_{ij}}{\epsilon_1}\right). \tag{14}$$

Combining the above expression with the marginal constraints $\boldsymbol{\pi}\mathbf{1} = \boldsymbol{a}$ and $\boldsymbol{\pi}^T\mathbf{1} = \boldsymbol{b}$, we obtain the following closed-form representation:

$$\exp\left(\frac{f_i}{\epsilon_1}\right) = \frac{a_i}{\sum_{j=1}^{K}\exp\left(\frac{g_j - C_{ij}}{\epsilon_1}\right)}. \tag{15}$$

$$\exp\left(\frac{g_j}{\epsilon_1}\right) = \frac{b_j}{\sum_{i=1}^{B}\exp\left(\frac{f_i - C_{ij}}{\epsilon_1}\right)}. \tag{16}$$

Next, we fix the remaining variables and update $\boldsymbol{b}$, that is,

$$\begin{aligned} &\frac{\partial L}{\partial b_j} = -\epsilon_2\left(\frac{1}{b_j} - \frac{1}{1-b_j}\right) + g_j - h = 0 \\ &\Leftrightarrow \ (g_j - h)b_j^2 + (h - g_j - 2\epsilon_2)b_j + \epsilon_2 = 0, \end{aligned} \tag{17}$$

whose discriminant is:

$$\Delta_j = (g_j - h)^2 + 4\epsilon_2^2 > 0. \tag{18}$$

We select the feasible root in $(0,1)$:

$$b_j = \frac{g_j - h + 2\epsilon_2 - \sqrt{\Delta_j}}{2(g_j - h)}. \tag{19}$$

Note that the other root $b_j = \frac{g_j - h + 2\epsilon_2 + \sqrt{\Delta_j}}{2(g_j - h)}$ falls outside the valid range $b_j \in [0,1]$ and is therefore discarded.

Then, using the normalization constraint $\boldsymbol{b}^T\mathbf{1} = 1$, substituting Eq. (19) yields:

$$f(h) = \boldsymbol{b}^T\mathbf{1} - 1 = \sum_{j=1}^{K}\frac{g_j - h + 2\epsilon_2 - \sqrt{\Delta_j}}{2(g_j - h)} - 1. \tag{20}$$

We use Newton's method to solve for the root of $f(h)$:

$$h \leftarrow h - \frac{f(h)}{f'(h)}. \tag{21}$$

Finally, by iteratively applying the update rules in Eqs.(15), (16), (21), and (19), we obtain the final transport matrix $\boldsymbol{\pi}$. Pseudo-labels for the current mini-batch are then assigned by selecting the cluster with the highest transport mass in each row of $\boldsymbol{\pi}$, yielding $\hat{\mathbf{Y}} = \{\hat{y}_i\}_{i=1}^{B}$ with entries defined as:

$$\hat{y}_{ij} = \begin{cases} 1, & \text{if } j = \arg\max_{j'} \pi_{ij'}, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

B Dataset Details

Following prior work, we conduct extensive experiments on six widely used short text clustering benchmark datasets. The statistical details of these datasets are summarized in Table 4, and a brief description of each dataset is provided below.

- **AgNews** (Zhang et al., 2015; Rakib et al., 2020): A benchmark of 8000 news headlines distributed evenly across four topical classes. Owing to its uniform label distribution, AgNews serves as a representative balanced short text dataset.

- **SearchSnippets** (Phan et al., 2008): A collection of 12340 short search engine snippets that span eight semantic domains. The category distribution is only mildly skewed, and the dataset is therefore regarded as lightly imbalanced, with texts that are sparse and fragmentary in nature.

- **GoogleNews** (Yin and Wang, 2016): Comprising 11109 news items associated with 152 events, this benchmark is provided in three forms that reflect different levels of textual granularity: GoogleNews-TS (titles and snippets), GoogleNews-T (titles only), and GoogleNews-S (snippets only). All three versions show large differences in class frequencies and therefore represent heavily imbalanced settings.

- **Tweet** (Yin and Wang, 2016): Consisting of 2472 microblog posts linked to 89 query topics from the TREC microblog tracks, this dataset is notably skewed, forming another heavily imbalanced testbed with highly informal and noisy text.

Dataset	*K*	*Size*	*Len*
AgNews (AN)	4	8000	23
SearchSnippets (SS)	8	12340	18
Googlenews-TS (G-TS)	152	11109	28
Googlenews-T (G-T)	152	11109	6
Googlenews-S (G-S)	152	11109	22
Tweet (TT)	89	2472	8

Table 4: The statistics of the six short text datasets. K denotes the number of clusters, *Size* denotes the number of short texts, and *Len* denotes the average text length.

C Baseline Details

To evaluate the performance of MAST, we compare it against a diverse set of representative short text clustering baselines, grouped into four categories: traditional frequency-based methods (BOW, TF-IDF), deep representation learning approaches (Self-Train, STCC, SBERT, BGE-M3), pseudo-label or semi-supervised optimization methods (RSTC, Multi-MCCR), and contrastive clustering models (SCCL, ProPos, CLSESSP, FNSCC). A brief introduction to these methods is provided below.

- **BOW and TF-IDF** (Bafna et al., 2016): These are traditional lexical feature extractors that convert documents into high-dimensional vectors. BOW encodes raw term counts, whereas TF-IDF reweights terms by their corpus-level distinctiveness to attenuate the impact of frequent but uninformative words.

- **Self-Train** (Hadifar et al., 2019): A deep clustering approach that refines sentence embeddings through an iterative self-training procedure. It combines autoencoder-based features with pre-trained sentence representations, then updates the encoder using cluster assignments as supervision to obtain more discriminative embeddings for short texts.

- **STCC** (Xu et al., 2017): A self-taught convolutional framework that learns deep short text representations in an unsupervised manner. It first obtains compact codes via unsupervised dimensionality reduction, then trains a CNN to fit these codes while leveraging word embeddings, and finally applies k-means on the learned representations.

- **SBERT** (Reimers and Gurevych, 2019): A Siamese-style extension of BERT that gener-

ates semantically meaningful sentence embeddings. By optimizing sentence-pair objectives, SBERT enables efficient similarity computation and provides strong baseline representations for clustering.

- **BGE-M3** (Xiao et al., 2024): A powerful multilingual sentence embedding model trained on large-scale curated corpora with enhanced contrastive objectives. BGE-M3 produces high-quality semantic representations across languages and has demonstrated strong generalization on standard embedding benchmarks, making it a solid baseline for short text clustering.

- **RSTC** (Zheng et al., 2023): A robust short text clustering framework designed to handle data imbalance and noise. RSTC consists of a pseudo-label generation module and a robust representation learning module. The pseudo-label module employs a self-adaptive OT formulation to produce balanced and noise-resistant pseudo-labels, which serve as supervision for representation refinement. The representation learning module integrates class-wise and instance-wise contrastive objectives to enhance cluster separability and mitigate the impact of noisy samples. RSTC achieves strong robustness across diverse datasets and is widely used as a competitive baseline in short text clustering.

- **Multi-MCCR** (Zhou et al., 2023): A semi-supervised text classification method that focuses on reducing the inconsistency introduced by dropout between training and inference. Multi-MCCR employs multiple parallel neural models with identical architectures to produce diverse output distributions for each input, enabling a richer basis for consistency regularization. It introduces a Consistent Bidirectional Kullback–Leibler divergence (C-BiKL) objective that jointly minimizes the bidirectional divergence among model predictions and the cross-entropy loss on labeled data. This design strengthens prediction consistency while preventing overfitting in low-resource settings. Although originally proposed for semi-supervised classification, Multi-MCCR serves as a strong baseline in clustering comparisons due to its stable optimization behavior and enhanced representation quality under limited supervision.

- **SCCL** (Zhang et al., 2021): A contrastive clustering framework designed to improve category separation in the representation space. SCCL enhances clustering by jointly leveraging instance-level contrastive learning and clustering-guided optimization, encouraging tighter intra-cluster cohesion and larger inter-cluster margins. It achieves substantial gains on short text clustering benchmarks, demonstrating the benefit of integrating contrastive discrimination with iterative cluster refinement.

- **ProPos** (Huang et al., 2022): A deep clustering method that integrates prototype scattering with positive sampling to balance representation uniformity and cluster compactness. ProPos enlarges distances between prototype embeddings to improve global separation, while aligning augmented instances with their sampled neighbors to enhance within-cluster coherence. By mitigating class-collision issues common in contrastive methods and avoiding collapse problems of non-contrastive approaches, ProPos achieves strong performance across both moderate-scale and large-scale clustering benchmarks.

- **CLSESSP** (Shen et al., 2024): A prototype-based contrastive learning framework that integrates semantic prototypes to address the limitations of traditional data augmentation in sentence embedding. CLSESSP constructs three prototype representations for each instance: a basic semantic prototype, a strong semantic prototype, and a negative semantic prototype, guided by prompt-based semantic cues. It optimizes a combination of the InfoNCE contrastive loss and a distribution divergence minimization objective to incorporate stronger semantic augmentation while preserving meaning. Experiments on semantic textual similarity datasets, transfer evaluations, and short text clustering benchmarks show that CLSESSP delivers consistent improvements over competitive baselines.

- **FNSCC** (Zheng et al., 2025b): A fuzzy neighborhood-aware contrastive clustering framework designed to address semantic sparsity and ambiguous boundaries in short texts.

FNSCC incorporates neighborhood information at both the instance and cluster levels. The instance-level module removes semantically close neighbors from the negative set to reduce false negatives and improve inter-cluster separability. The cluster-level module introduces fuzzy neighborhood-aware weighting to refine soft assignment probabilities and encourage alignment with semantically coherent groups. By jointly modeling local semantic structure and cluster-level relationships, FNSCC enhances representation quality and consistently delivers strong performance on multiple short text clustering benchmarks.

D Evaluation Metrics

Following prior studies on short text clustering, we evaluate model performance using two widely adopted metrics: Accuracy (ACC) and Normalized Mutual Information (NMI). ACC quantifies the alignment between predicted clusters and gold labels by identifying the optimal one-to-one mapping between them. NMI assesses the similarity between the predicted and true label distributions by normalizing their mutual information with respect to their entropies. Both metrics take values in the range $[0, 1]$, with higher scores indicating stronger clustering consistency and overall quality.

ACC evaluates the proportion of correctly matched cluster assignments using an optimal permutation ϕ (Hungarian mapping function) between predicted and true labels:

$$\text{ACC} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\Big(\ell_j = \phi(\tilde{\ell}_j)\Big), \tag{23}$$

where ℓ_j and $\tilde{\ell}_j$ denote the ground-truth and predicted labels for instance j, respectively.

NMI quantifies the agreement between predicted clusters $\tilde{L}$ and true labels L through their mutual dependence:

$$\text{NMI} = \frac{I(L, \tilde{L})}{\sqrt{H(L)\, H(\tilde{L})}}, \tag{24}$$

where $I(L, \tilde{L})$ denotes mutual information and $H(\cdot)$ is the entropy of a label distribution.

E Iterative Visualization on SearchSnippets

We visualize MAST on SearchSnippets using t-SNE in Figure 5. Colors denote predicted clusters and black dots denote cluster centers. Early in training, embeddings exhibit substantial overlap with weak cluster structure. As training proceeds, representations separate and form compact groups, yielding clearer cluster boundaries and a more coherent global geometry.

F Graph Encoder Variants for Structural Views

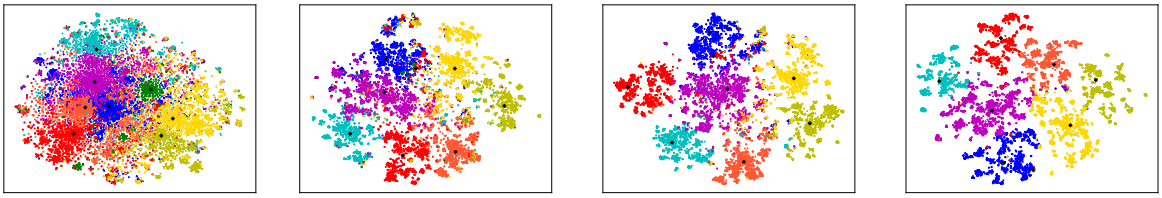
To investigate the effect of different graph encoders used for structural view construction in MAST, we replace the default GCN (Kipf and Welling, 2017) with two widely used graph neural network variants, namely GraphSAGE (Hamilton et al., 2017) and GAT (Velickovic et al., 2018). All other components and hyperparameters are kept identical to ensure a fair comparison. Table 5 reports the clustering performance on six benchmark datasets in terms of ACC and NMI.

As shown in Table 5, GCN consistently achieves the best or near-best performance across all datasets. GraphSAGE and GAT remain competitive but generally underperform GCN, with larger gaps observed on datasets exhibiting higher semantic sparsity or class imbalance. GCN benefits from a normalized aggregation scheme that provides a stable inductive bias, effectively smoothing noisy relations while preserving core semantic structure. GraphSAGE introduces additional sampling and aggregation overhead, which may lead to information loss and unstable estimates in small or noisy sentence graphs. GAT further increases computational complexity by learning attention weights for each edge, amplifying sensitivity to noisy similarities in unsupervised short text clustering.

Importantly, these results also demonstrate the flexibility of MAST with respect to graph encoder choices. MAST remains effective across different graph architectures, indicating that the proposed multi-view alignment and clustering optimization framework generalizes well to diverse structural modeling strategies. Given its favorable balance between performance, efficiency, and stability, GCN is adopted as the default graph encoder in MAST.

G Case Study and Error Analysis

In addition to the clustering-evolution visualization on the SearchSnippets dataset shown in Figure 5, we further examine the fine-grained clustering capability of MAST on real short texts. We select a representative subset of samples from Search-

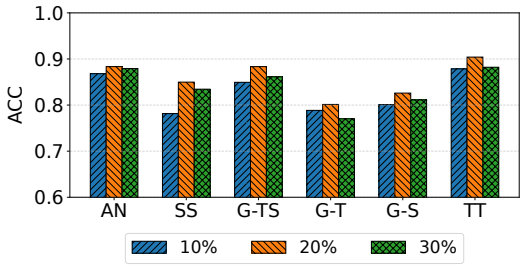


(a) Raw (b) Iteration 500 (c) Iteration 1000 (d) Final

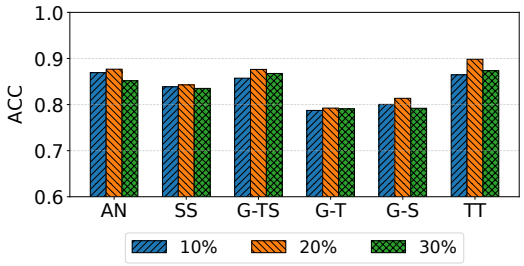
Figure 5: Clustering visualization results for the SearchSnippets text dataset. Different colors denote distinct clusters and black dots indicate the corresponding cluster centers.

	AN		SS		G-TS		G-T		G-S		TT	
Settings	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
MAST-GraphSAGE	88.13	67.93	83.88	68.74	87.83	94.25	78.97	89.45	81.23	89.71	89.78	93.58
MAST-GAT	**88.56**	**68.64**	84.23	69.22	**88.52**	94.48	78.62	89.23	81.42	89.77	89.64	93.55
MAST-GCN(ours)	88.34	68.39	**84.98**	**69.94**	88.34	**94.57**	**80.14**	**89.65**	**82.60**	**89.81**	**90.41**	**93.87**

Table 5: Performance (%) comparison of MAST with different graph encoders for structural view construction. GCN denotes the default setting used in MAST. **Bold** indicates a better performance.



(a) BERT



(b) RoBERTa

Figure 6: ACC of MAST under different word replacement ratios, where the augmented-view is generated by BERT or RoBERTa.

Snippets that are prone to semantic ambiguity near cluster boundaries and compare their clustering results across different ablation variants of MAST. As illustrated in Figure 7, these texts contain cross-domain terms such as *quantum*, *entertainment*, and *politics*. These terms may appear in business and finance contexts while remaining semantically close to topics such as Politics, Entertainment, and Sports News. This overlap introduces ambiguity and causes partially optimized models to misassign samples across related clusters.

When comparing the clustering outcomes of different variants, we observe the following patterns:

(1) w/o CAO. Removing the contrastive alignment objective reduces the model's ability to capture subtle semantic differences. Several samples are incorrectly assigned to unrelated clusters (e.g., XML Technology or Sports News). This observation shows that maintaining consistency across views is essential for forming a clearly separated semantic space.

(2) w/o HGO. Removing high-confidence guided optimization weakens the stability of cluster center updates. The model correctly classifies only part of the samples, and the remaining ones oscillate between adjacent clusters. This result indicates that HGO plays an important role in reinforcing cluster structures and suppressing noise-induced drift.

(3) w/o OTPA. Removing OT-based pseudo-label alignment preserves a certain degree of discriminative ability but cannot fully correct the bias introduced by augmented-views. Some samples are misassigned to the Entertainment News cluster, demonstrating that OTPA contributes significantly to aligning cross-view pseudo-label distributions.

Compared with the above variants, the complete MAST framework, which combines multi-view contrastive alignment with clustering optimization, produces stable and correct predictions for all samples. It effectively eliminates category drift caused

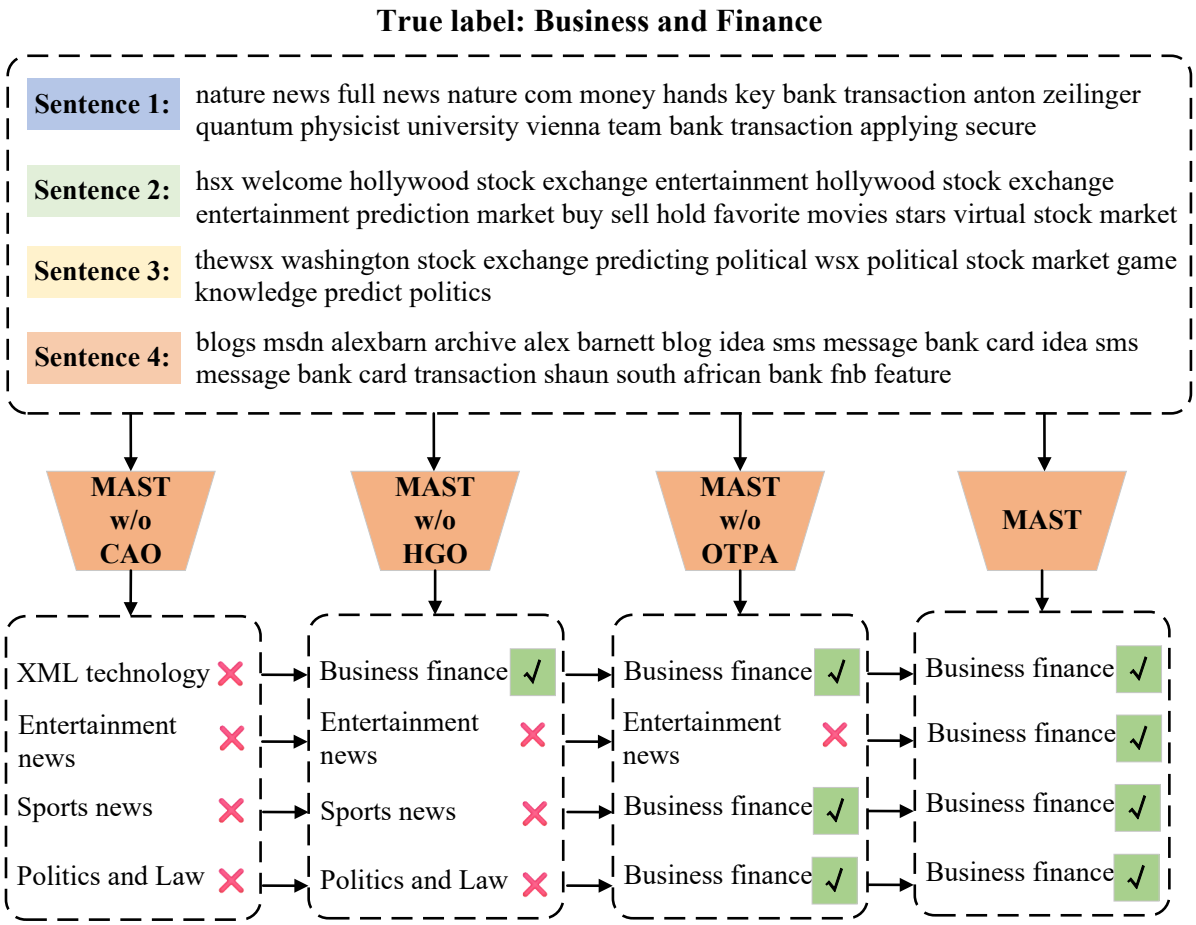


Figure 7: A case study on SearchSnippets with true label Business and Finance.

by fuzzy boundaries. This case study further verifies the complementary strengths of multi-view alignment, high-confidence optimization, and OT-based pseudo-label correction, and shows that they jointly lead to a more robust clustering structure for short text data.

H Data Augmentation

Existing studies (Zhang et al., 2021; Zheng et al., 2025b) show that the effectiveness of contrastive learning in clustering tasks strongly depends on the quality of the augmented texts. Among various augmentation strategies, the Contextual Augmenter performs particularly well for short text scenarios. It leverages a masked language model to perform token-level substitution, producing augmented-views with richer semantics and more complete contextual cues. This helps mitigate the inherent semantic sparsity of short texts and improves the discriminative power of contrastive objectives.

Motivated by this, we adopt the Contextual Augmenter in our experiments and systematically evaluate how different masked language models (BERT and RoBERTa) and substitution rates affect the performance of MAST. As illustrated in Figure 6, BERT-based augmentation with a 20% substitution rate yields the most significant performance gain. Consequently, we use this setting as the default configuration in our main experiments.

I Discussion on Training Time

All experiments are conducted on a single GeForce RTX 4090 GPU. The training time of MAST on the six benchmark datasets ranges from roughly 10 to 20 minutes per dataset, which is comparable to representative contrastive clustering methods such as RSTC (Zheng et al., 2023) and FNSCC (Zheng et al., 2025b). Despite incorporating multi-view structural modeling and OT-based refinement, the overall training cost remains practical.

We further analyze the computational complexity of MAST. The most time-consuming components of MAST are: (i) dual-graph propagation on the sentence-level graph $\mathcal{G}_S$ and the cluster-level graph $\mathcal{G}_C$; (ii) the multi-view contrastive alignment with structure-aware negatives and intermediate-

layer negatives; and (iii) the OT-based pseudo-label refinement.

For two-layer GCNs, the costs on $\mathcal{G}_S$ and $\mathcal{G}_C$ are $O(E_S(d_S+d)+Bd^2)$ and $O(E_C(d_C+d)+Kd^2)$, respectively, where B is the batch size, K is the number of clusters, d_S and d_C are input dimensions, d is the hidden dimension, and E_S, E_C denote the numbers of non-zero edges in $\mathbf{A}_S$ and $\mathbf{A}_C$. The multi-view contrastive objective requires $O(B^2d + Bd)$ operations per batch, since all in-batch pairs across a constant number of views are compared in the shared embedding space. The OT module operates on a $B \times K$ cost matrix with L Sinkhorn iterations, leading to an additional cost of $O(LBK)$.

Overall, the time complexity of MAST can be summarized as $O\big(B^2d+(E_S+E_C)d+(B+K)d^2+LBK\big)$, which is comparable to existing contrastive clustering frameworks in practice.