FNSCC: Fuzzy Neighborhood-Aware Self-Supervised Contrastive Clustering for Short Text

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

Abstract

Short texts pose significant challenges for clustering due to semantic sparsity, limited context, and fuzzy category boundaries. Although recent contrastive learning methods improve instance-level representation, they often overlook local semantic structure within the clustering head. Moreover, treating semantically similar neighbors as negatives impair cluster-level discrimination. To address these issues, we propose Fuzzy Neighborhood-Aware Self-Supervised Contrastive Clustering (FNSCC) framework. FNSCC incorporates neighborhood information at both the instance-level and cluster-level. At the instance-level, it excludes neighbors from the negative sample set to enhance inter-cluster separability. At the cluster-level, it introduces fuzzy neighborhood-aware weighting to refine soft assignment probabilities, encouraging alignment with semantically coherent clusters. Experiments on multiple benchmark short text datasets demonstrate that FNSCC consistently outperforms state-of-the-art models in accuracy and normalized mutual information. Our code is available at https://github.com/zjzone/FNSCC.

1 Introduction

Text clustering, a fundamental task in natural language processing, seeks to uncover the latent semantic structure of text without relying on prior annotations, and provides theoretical support for downstream applications (Wan et al., 2024; Wu et al., 2020). Its significance is particularly evident in domains such as information retrieval (Zhao et al., 2022), knowledge discovery (Guan et al., 2020), and natural language understanding (Saharia et al., 2022).

Text clustering is inherently challenging due to high-dimensional sparsity, semantic ambiguity, and class imbalance. Effective clustering requires capturing multi-level semantic features. Traditional methods such as Bag-of-Words (BOW) and TF-IDF rely on word frequency and lack contextual awareness. Word2Vec improves semantic representation by learning context-aware embeddings through predictive objectives and efficient training strategies like negative sampling. BERT (Kenton and Toutanova, 2019) and other pre-trained language models (PLMs) leverage global attention and positional encoding to capture long-range dependencies. However, prior work (Ethayarajh, 2019) has shown that their sentence representations perform poorly in clustering tasks. This limitation stems from the fact that BERT is pre-trained with objectives such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), which are not designed to capture global semantic categories or learn clustering-friendly representations. To overcome this, SCCL (Zhang et al., 2021) proposes a deep clustering framework that fine-tunes PLMs with joint contrastive and clustering objectives, improving semantic separability and explicitly modeling cluster structures for better clustering performance.

Since SCCL aims to generate high-confidence sample partitions that align with the underlying clustering structure. However, it relies solely on the Euclidean distance between text instance embeddings and cluster centers to compute soft cluster assignments, which can be susceptible to misjudgments caused by fuzzy noise. This limitation prevents high-confidence samples from receiving sufficient emphasis during training. Additionally, in the contrastive learning module, neighboring samples often have higher semantic correlation. If only the corresponding augmented instances are taken as positive examples, these semantically similar samples are considered negative samples, and the model force the distance between them to be increased, thereby destroying the clustering structure.

To address the aforementioned issues, we

* Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2831–2846

propose the Fuzzy Neighborhood-Aware Self-Supervised Contrastive Clustering (FNSCC) framework for short text. The primary goal is to enhance the clustering structure and instance contrastive module of SCCL by incorporating fuzzy neighborhood information. Specifically, we first construct a more balanced distribution of positive and negative samples in the instance contrastive module to prevent semantically similar samples from being classified as negative examples. Additionally, we leverage the neighborhood semantic information of the current sample to improve the expressiveness of the local context in the clustering head through a fuzzy weighting mechanism, thereby mitigating the impact of single-point noise on soft clustering.

In summary, the main contributions of this paper are as follows:

(1) Neighborhood-aware negative sampling improves contrastive learning by excluding neighboring instances from the negative sample set. Such a refinement reduces semantic interference from similar samples and enhances the model's ability to distinguish positives from hard negatives.

(2) Fuzzy neighborhood-based assignment weighting introduces local structural awareness into the clustering head, allowing for more coherent cluster boundaries and higher confidence in soft assignment probabilities.

(3) We propose an end-to-end framework named FNSCC that jointly optimizes neighborhood-aware contrastive learning and clustering. Extensive experiments on multiple benchmark datasets demonstrate that FNSCC consistently outperforms existing methods on short text clustering tasks. Ablation studies further confirm the effectiveness of each component.

2 Related Works

2.1 Text Clustering

Text clustering has developed through three main stages. Traditional methods, such as the BOW model and TF-IDF, are simple and effective for small datasets but lack semantic awareness and struggle with high-dimensional sparse data (Shi et al., 2024). The second stage uses shallow neural networks to convert sparse word frequency vectors into dense semantic embeddings, capturing contextual relationships and semantic similarity (Lin and Lin, 2023). The third stage introduces deep clustering methods, which combine clustering with PLMs or various feature learning strategies to enhance optimization (Cai et al., 2022; Gupta et al., 2022). However, these methods often prioritize feature representation, which limits their clustering effectiveness.

Recent studies have explored graph-based representations for feature learning, yielding promising results (Huang et al., 2021; Hua et al., 2023). For example, Chiu et al. (2020) construct keyword correlation graphs and leverage graph autoencoders to capture local and global document features, though their method lacks clustering-specific optimization. Zhang et al. (2021) achieve state-of-the-art performance by combining clustering and contrastive learning within the SBERT framework (Reimers, 2019). Building on this, our approach incorporates fuzzy neighborhood information into the SCCL framework to better capture local data distributions. This refinement enhances clustering stability and contrastive learning effectiveness, addressing key limitations of prior methods.

2.2 Self-Supervised Learning

Self-supervised learning (SSL) (Ermolov et al., 2021; Baevski et al., 2022) has emerged as a powerful paradigm in representation learning, offering an efficient alternative to traditional unsupervised methods. By designing pretext tasks that derive optimization objectives directly from data, SSL extracts meaningful and transferable semantic representations for downstream tasks. Traditional SSL methods based on autoencoders and generative models (Eckart et al., 2021; Hou et al., 2022) effectively capture feature distributions but often face limited generalization due to their task-specific nature. In contrast, contrastive learning gains popularity by enhancing semantic separation through aligning augmented samples while repelling dissimilar ones (Cui et al., 2021; Xu et al., 2022). Recent improvements incorporate neighborhood information to further boost representation quality in both text and vision domains (Zhong et al., 2021; Sun et al., 2023).

In the SCCL framework, cluster assignments are computed using Student's t-distribution (Xie et al., 2016), and then refined into target distributions for self-supervised clustering. Building on this, we propose a neighborhood-aware self-supervised loss that integrates fuzzy neighborhood information into both the clustering and contrastive modules. These improvements enhance clustering stability and representation quality, addressing key limitations of prior work.

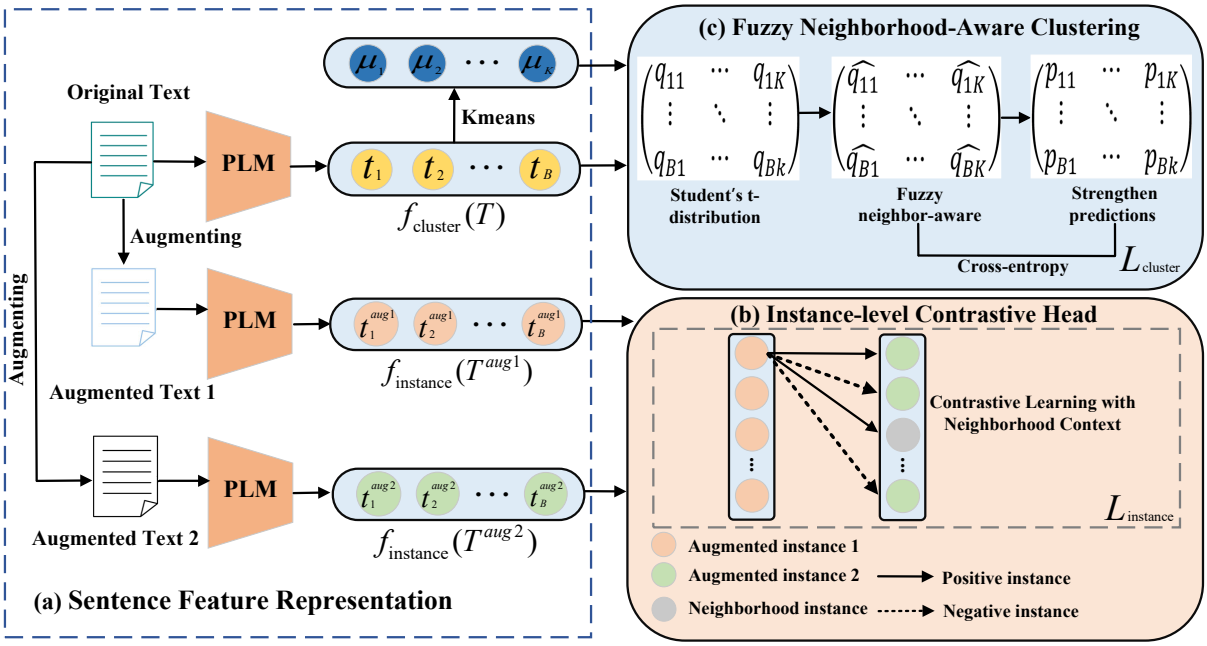


Figure 1: Overview of the FNSCC framework. Given input text T and its augmented views T^{aug1} and T^{aug2}, a shared encoder produces sentence embeddings $f_{cluster}(T)$, $f_{instance}(T^{\text{aug1}})$, and $f_{instance}(T^{\text{aug2}})$. The clustering head refines soft assignments on $f_{cluster}(T)$ using high-confidence fuzzy neighborhoods, while the instance-level contrastive head leverages $f_{instance}(T^{\text{aug1}})$ and $f_{instance}(T^{\text{aug2}})$, excluding neighbors of the anchor sample from the negative set to improve representation discrimination.

3 The Proposed FNSCC Method

In this section, we provide a detailed description of the FNSCC method. Section 3.1 outlines the general framework and clustering objectives of FNSCC. Sections 3.2-3.3 elaborate on the primary contributions of this work to contrastive clustering, specifically the fuzzy neighborhood-enhanced clustering mechanism, which includes both instance-level and cluster-level modules.

3.1 General Framework of the Contrastive Clustering

To improve clustering confidence and contrastive learning, FNSCC employs a neighborhood-aware design. In the representation stage, a batch of original instances $T = \{t_1, ..., t_i, ..., t_B\}$ is sampled, and two augmented views are generated using a contextual augmenter (Kobayashi, 2018): $T^{aug1} = \{t_1^{aug1}, ..., t_i^{aug1}, ..., t_B^{aug1}\}$ and $T^{aug2} = \{t_1^{aug2}, ..., t_i^{aug2}, ..., t_B^{aug2}\}$. This strategy enriches short text information and supports semantic learning from multiple perspectives, with details provided in Appendix A.7.

FNSCC uses a pre-trained SBERT (Reimers, 2019) to obtain original text embeddings and applies an MLP to refine augmented embeddings, enhancing diversity and reducing redundancy. The representations are projected into d_1- and d_2-dimensional spaces: $f_{\text{cluster}}(T) = \{f_{\text{cluster}}(t_i) \in \mathbb{R}^{d_1}\}_{i=1}^{B}$ and $f_{\text{instance}}(T^{\text{aug}}) = \{f_{\text{instance}}(t_i^{\text{aug}}) \in \mathbb{R}^{d_2}\}_{i=1}^{B}$, using the encoders f_{cluster} and f_{instance}, respectively.

The design serves two main purposes: (1) $f_{cluster}(T)$ is optimized to enhance the expressive power of the clustering head, making the embeddings of the original texts more discriminative in clustering tasks, and (2) $f_{instance}(T^{aug})$ is optimized through instance-level contrastive learning, improving the robustness and generalization of text embeddings for distinguishing between categories.

Our innovation lies in the objective function $\mathcal{L}$, which combines instance contrastive loss and clustering loss for joint optimization. The model is trained end-to-end to produce effective cluster-friendly representations. The overall objective function $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{instance} + \beta \cdot \mathcal{L}_{cluster}, \qquad (1)$$

where $\mathcal{L}_{instance}$ denotes the instance-level contrastive loss, designed to enhance similarity learning between instances, while $\mathcal{L}_{cluster}$ represents the cluster-level loss, which optimizes the clustering head by ensuring that samples within the same cluster exhibit greater similarity. The hyperparam-

eters α and β are introduced to balance $\mathcal{L}_{instance}$ and $\mathcal{L}_{cluster}$, they are set to 10 and 1 (1 and 1 for the Tweet dataset), respectively. Following the aforementioned steps, the flowchart of FNSCC is illustrated in Figure 1.

Finally, the k-means clustering algorithm is applied to the text feature vectors learned by the model to obtain the final clustering results.

3.2 Instance-Level Contrastive Learning with Neighborhood Context

In $\mathcal{L}_{instance}$, the optimization objectives include two key aspects: (1) maximizing the similarity between two augmented instances (positive instances) to ensure that the two versions generated from the same instance remain consistent in the feature space, thereby enhancing the cohesion of objects within the same cluster; (2) minimizing the similarity between non-neighbor negative instances, i.e., reducing the similarity between augmented instances and other non-neighbor instances outside the cluster, effectively increasing the separation between different clusters. This contrastive strategy facilitates the construction of a cluster-friendly representation space, ensuring tight alignment of similar instances and clear separation of dissimilar ones.

We define neighboring instances using a k-nearest neighbors (KNN) approach based on cosine similarity between instance embeddings. For each instance, the top-k most similar samples in the embedding space are selected as its neighbors. By explicitly modeling such local structures, we aim to preserve intra-cluster compactness and mitigate the potential impact of noisy negative samples in contrastive learning.

Since neighboring instances may belong to the same cluster, treating them as negative instances during training could introduce noise and impair the model's ability to distinguish between positive and negative instances. To address this, we exclude neighboring samples when calculating negative examples, avoiding their misclassification as negatives. This approach enables the model to leverage local similarity information more effectively during feature learning, enhancing clustering-specific characteristics and ultimately improving clustering performance.

Let $E_i^{aug} = f_{instance}(t_i^{aug})$ denote the representation of the augmented text. The loss function for augmented instances t_i^{aug1} is defined as follows:

$$\ell(t_i^{\text{aug1}}) = -\log\left[\exp\left(s\left(E_i^{\text{aug1}}, E_i^{\text{aug2}}\right)/\tau\right) / \sum_{j=1}^{2B} \mathbb{1}_{\{j \notin \mathcal{N}(E_i^{\text{aug1}}) \cup \{i\}\}} \exp\left(s\left(E_i^{\text{aug1}}, E_j^{(\text{aug1,aug2})}\right)/\tau\right)\right], \tag{2}$$

where B denotes the number of instances in the current batch, $s(\cdot)$ represents the cosine similarity, $\mathcal{N}(\cdot)$ represents the neighborhood index corresponding to the instance, meaning that the k instances with the largest cosine similarity to the current instance are regarded as neighboring instances. $\mathbb{1}$ is an indicator function and τ is the temperature coefficient, which is set to 0.5. The contrastive loss for the augmented instances, including the entire batch of B instances, is defined as follows:

$$\mathcal{L}_{\text{instance}} = \frac{1}{2B}\sum_{i=1}^{B} \ell(t_i^{\text{aug1}}) + \ell(t_i^{\text{aug2}}). \tag{3}$$

3.3 Fuzzy Neighborhood-Aware Clustering

Unlike instance-level contrastive learning, the clustering head aims to group instances into the same cluster, enabling the model to capture higher-level semantic commonalities rather than individual distinctions. To emphasize local features and balance the influence of center and boundary samples, we adopt the method from Xie et al. (2016) to compute the soft assignment of an instance $t_i \in T$ to cluster center μ_k, where $k \in \{1, 2, \ldots, K\}$ and K denotes the number of clusters. Let $g(t_i) = f_{\text{cluster}}(t_i)$ be the cluster representation of instance t_i. The probability q_{ik}, indicating the likelihood that $g(t_i)$ is assigned to μ_k, is defined as:

$$q_{ik} = \frac{(1 + \|g(t_i) - \mu_k\|_2^2/\gamma)^{-\frac{\gamma+1}{2}}}{\sum_{k'=1}^{K}(1 + \|g(t_i) - \mu_{k'}\|_2^2/\gamma)^{-\frac{\gamma+1}{2}}}, \tag{4}$$

where γ represents the degrees of freedom of the Student's t-distribution. Consistent with the method in SCCL, we set γ to 1.

While this method provides a soft probability distribution over cluster assignments, it does not consider the fact that neighboring samples often belong to the same cluster. To incorporate this structural prior, we propose a fuzzy neighborhood-aware refinement of the cluster assignment. In this context, the term "fuzzy" does not refer to fuzzy logic but rather to the soft incorporation of neighborhood information into the probability distribution. The updated assignment probability q_{ik}^* is

computed as follows:

$$q_{ik}^{*} = \frac{q_{ik} + \lambda \sum_{j \in \mathcal{N}(g(t_i))} \Psi(t_i, t_j) \cdot q_{jk}}{1 + \lambda \sum_{j \in \mathcal{N}(g(t_i))} \Psi(t_i, t_j)}, \quad (5)$$

where $\Psi(t_i, t_j) = s(g(t_i), g(t_j)) \cdot d(g(t_i), g(t_j))$. As in Equation (2), $\mathcal{N}(\cdot)$ denotes the neighbor index set for a given text instance. This definition is consistent with that used in the instance contrastive head, ensuring semantic alignment within the local neighborhood. λ is the neighbor weight coefficient, set to 0.5. The function $s(\cdot)$ represents cosine similarity, capturing semantic closeness between instances. The function $d(\cdot)$ models neighborhood density using a Gaussian kernel based on Euclidean distance, assigning higher weights to neighbors that are closer.

In Equation (4), cluster membership relies solely on Euclidean distance and thus ignores the local data structure. To address this limitation, we introduce the fuzzy neighborhood weighting factor $\Psi(t_i, t_j)$, which measures the influence of neighbor t_j on the soft assignment of instance t_i. By combining semantic similarity and local density, this factor enables the model to preserve the global clustering structure while fully leveraging local relationships. Compared to traditional methods based only on distance, this approach enhances the robustness and accuracy of clustering assignments.

To ensure interpretability and probabilistic consistency, the assignment scores for each instance are further normalized to sum to 1. This yields a well-calibrated probability distribution for downstream clustering. The final allocation probability $\hat{q}_{ik}$ is defined as follows:

$$\hat{q}_{ik} = \frac{q_{ik}^{*}}{\sum_k q_{ik}^{*}}. \quad (6)$$

We then introduce an auxiliary target distribution p, derived from the Student's t-distribution, to emphasize high-confidence instances and further optimize cluster centers. The auxiliary probability p_{ik} is defined as follows:

$$p_{ik} = \frac{\hat{q}_{ik}^{2} / \sum_{j=1}^{B} \hat{q}_{jk}}{\sum_{k'} (\hat{q}_{ik'}^{2} / \sum_{j=1}^{B} \hat{q}_{jk'})}. \quad (7)$$

The main idea is to assign greater weights to high-confidence samples while down-weighting uncertain ones, enabling gradual optimization of cluster centers. To align soft assignments with the target distribution, we adopt a cross-entropy loss for the clustering head.

$$\mathcal{L}_i = -\sum_{k=1}^{K} p_{ik} \log q_{ik}. \quad (8)$$

We then define the clustering objective for each mini-batch of size B as follows:

$$\mathcal{L}_{cluster} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_i. \quad (9)$$

By optimizing the model with the loss from the clustering head, we leverage the benefits of the target distribution to guide the model in producing more accurate and stable clustering results.

The detailed parameter analysis of the model will be further discussed in Appendix A.4.

4 Experiments

In this section, we conduct experiments on several real-world short text datasets to demonstrate the contribution of our method.

4.1 Experimental Setup

4.1.1 Datasets

The FNSCC algorithm is evaluated on six widely used benchmark short text datasets: AgNews, SearchSnippets, GoogleNews-TS, GoogleNews-T, GoogleNews-S, and Tweet. Descriptions and statistics for these datasets are provided in Appendix A.1.

4.1.2 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 et al., 2016). **(II)** Representation learning-based methods include **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 include **SCCL** (Zhang et al., 2021), **ProPos** (Huang et al., 2022) and **CLSESSP** (Shen et al., 2024). **(IV)** Semi-supervised and pseudo-label optimization-based methods include **Multi-MCCR** (Zhou et al., 2023) and **RSTC** (Zheng et al., 2023). Descriptions of these methods are provided in Appendix A.2.

	AgNews		SearchSnippets		GoogleNews-TS	
Model	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
BOW	27.60	2.60	24.30	9.30	57.50	81.90
TF-IDF	34.50	11.90	31.50	19.20	68.00	88.90
STCC	83.50	56.90	77.00	56.60	76.90	80.60
Self-Train	63.60	35.50	77.10	56.70	59.40	79.60
SBERT(k-means)	83.44	57.76	73.02	59.77	67.40	90.47
BGE-M3	87.59	-	80.57	-	56.28	-
SCCL	84.62	62.73	75.86	63.67	79.24	92.31
ProPos	84.30	59.30	74.30	55.20	73.90	90.40
CLSESSP	80.45	-	69.85	-	64.53	-
Multi-MCCR	87.10	-	80.59	-	51.42	-
RSTC	85.98	64.32	79.75	**69.48**	79.93	92.60
FNSCC	**87.85**	**66.70**	**82.59**	67.65	**88.21**	**94.31**

	GoogleNews-T		GoogleNews-S		Tweet	
Model	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
BOW	49.80	73.20	49.00	73.50	49.70	73.60
TF-IDF	58.90	79.30	61.90	83.00	57.00	80.70
SBERT(k-means)	63.98	86.13	65.87	87.64	62.70	86.80
BGE-M3	49.88	-	52.07	-	77.66	-
SCCL	67.32	84.73	78.94	**89.37**	75.49	89.06
ProPos	65.41	85.32	75.57	87.19	78.42	88.53
CLSESSP	63.60	-	64.64	-	57.85	-
Multi-MCCR	43.33	-	47.32	-	72.34	-
RSTC	**75.50**	**88.39**	76.01	88.27	75.20	85.62
FNSCC	72.72	87.76	**80.49**	**89.37**	**83.62**	**90.38**

Table 1: Clustering performance comparison on six real short text datasets. We highlight the best performance in bold.

4.1.3 Implementation

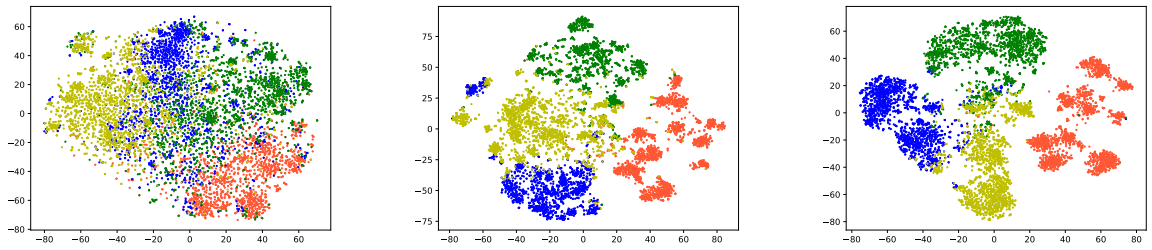
The FNSCC method is implemented in PyTorch (Paszke et al., 2019), using Sentence Transformer (Reimers, 2019) as the backbone for the PLM.

The tokenizer's maximum input length is set to 32. To enhance the instance contrastive module, an MLP is introduced, consisting of a single hidden layer and an output layer, both with a vector size of 768. The learning rate for the baseline model is set to 1e-5, while the learning rates for the clustering head and instance contrastive head are set to 1e-3. The Adam optimizer is used with a batch size of 128, and training runs for 3000 iterations. Furthermore, the datasets are augmented using BERT and RoBERTa (Liu et al., 2019) with a 20% word substitution rate. To objectively evaluate the proposed method, we employ accuracy (ACC) and normalized mutual information (NMI) as evaluation metrics, consistent with those used in the comparison algorithms. Details of these metrics are provided in Appendix A.3. Considering the inherent randomness in the training process and the instability of k-means clustering, the results presented in this paper represent the average of five experiments.

4.2 Main Results

We evaluate FNSCC against several representative text clustering models on six benchmark datasets. Table 1 reports the results. FNSCC consistently achieves state-of-the-art performance on most datasets. While RSTC slightly surpasses FNSCC on SearchSnippets (NMI) and GoogleNews-T (ACC and NMI), the overall results validate the effectiveness of integrating fuzzy neighborhood information into the clustering module and applying instance-level contrastive loss. These improvements enhance both clustering stability and accuracy, contributing to the superior performance of FNSCC.

FNSCC demonstrates strong performance on



(a) SBERT(k-means) (b) SCCL (c) FNSCC

Figure 2: Clustering visualization results for the AgNews text dataset.

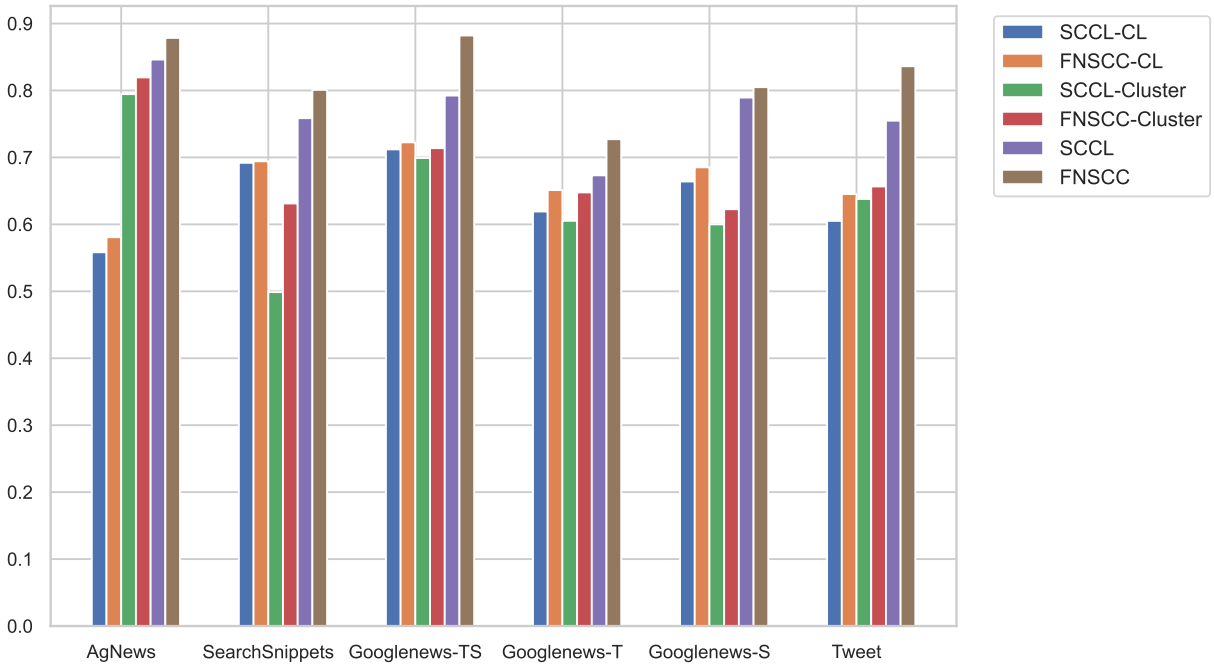


Figure 3: Ablation results of FNSCC in terms of ACC under configurations that exclude in-neighborhood false negatives at the instance-level and eliminate fuzzy neighborhood-based confidence estimation at the cluster-level.

datasets with a small number of clusters, such as AgNews and SearchSnippets. SearchSnippets presents unique challenges, as it mainly consists of isolated keywords or short phrases rather than complete sentences. This structure leads to sparse semantics, imbalanced class distributions, and weak contextual cues, limiting conventional neighborhood modeling and increasing encoder ambiguity. By emphasizing high-confidence neighborhood semantics, FNSCC improves local semantic discrimination and achieves the highest clustering accuracy (ACC). However, its normalized mutual information (NMI) is slightly lower than RSTC's, due to the dataset's restricted global structure and NMI's sensitivity to global label consistency. On GoogleNews-T, which contains short titles with limited context, FNSCC performs comparably to RSTC in both ACC and NMI. RSTC benefits from using global pseudo-labels to maintain label alignment across instances. In contrast, GoogleNews-S and GoogleNews-TS offer more contextual richness through snippets or combined formats, enabling FNSCC to better exploit neighborhood information. As a result, FNSCC achieves strong clustering results on these datasets, highlighting its robustness across different text granularities.

Additionally, we re-implement the SCCL model (Zhang et al., 2021) with the two masked language models used here, following the original parameter settings, to compare it with FNSCC.

To compare SCCL and FNSCC and to illustrate the effects of instance-level contrastive learning with neighborhood context and fuzzy neighborhood-aware clustering on representation

quality, we utilize the t-SNE technique (Fujiwara et al., 2021) to project the high-dimensional clustering distributions into two dimensions for visualization after text embedding. As shown in Figure 2, the results indicate that, compared to the unoptimized PLM SBERT and SCCL without fuzzy neighborhood information, FNSCC achieves greater effectiveness in distinguishing clusters and ensuring cohesion within the same cluster.

Additional experiments are provided in the appendices: (1) training loss curves and clustering performance (Appendix A.5); (2) ablation studies assessing the contribution of each FNSCC component via NMI (Appendix A.6); (3) evaluation of data augmentation strategies (Appendix A.7); and (4) analysis of failure cases (Appendix A.8).

4.3 Ablation Study

To investigate how fuzzy neighborhoods influence clustering performance, we systematically ablate and reweight the components of FNSCC, analyzing their individual contributions to the model's behavior.

We evaluate six variants: SCCL-CL, which applies instance-level contrastive learning without neighbor filtering; SCCL-Cluster, which applies cluster-level optimization without fuzzy neighborhood enhancement; FNSCC-CL, which incorporates neighborhood-aware instance-level contrastive learning by removing in-neighborhood samples from negative sampling; FNSCC-Cluster, which introduces fuzzy neighborhood-aware assignment smoothing at the cluster level; and the full models SCCL and FNSCC. Clustering results measured by ACC are shown in Figure 3.

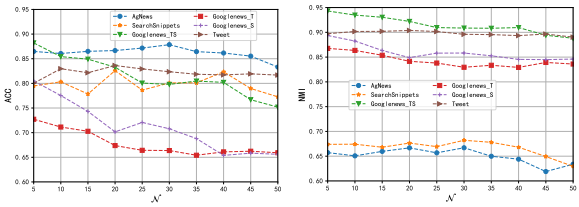
As shown in the figure, both instance-level and cluster-level modules benefit significantly from incorporating fuzzy neighborhood mechanisms. Specifically, FNSCC-CL outperforms SCCL-CL, demonstrating the effectiveness of excluding nearby neighbors from negative samples, which prevents semantically similar instances from being pushed apart. This encourages more discriminative yet semantically coherent instance representations. At the cluster level, FNSCC-Cluster outperforms SCCL-Cluster by integrating fuzzy neighborhood information into soft assignment, promoting local semantic consistency and improving confidence in cluster boundaries.

Furthermore, the full FNSCC model achieves the best performance, confirming that jointly optimizing both levels with neighborhood awareness yields more robust clustering. Similar trends are observed under the NMI metric, as reported in Appendix A.6.

4.4 Neighborhood Sensitivity Analysis

Given that the instance contrastive head and the clustering head are aligned in terms of neighborhood structures, both modules in the proposed method effectively utilize the setting of the hyperparameter $\mathcal{N}$. The performance improvement of FNSCC is largely dependent on the auxiliary role played by the $\mathcal{N}$ neighborhood. Therefore, it is crucial to assess the sensitivity of FNSCC to the size of $\mathcal{N}$.

This section examines the performance of FNSCC across different datasets by adjusting the $\mathcal{N}$ parameter and provides recommendations based on the findings. Figure 4 presents the ACC and NMI values for different values of $\mathcal{N}$ on six short text datasets. The parameter $\mathcal{N}$ is varied within the range of 5 to 50. As shown in Figure 4, both ACC and NMI exhibit fluctuations within a certain range as $\mathcal{N}$ increases, with NMI demonstrating relatively more stability. It can be observed that performance improves gradually as $\mathcal{N}$ increases from 5 to 30 but begins to decline when $\mathcal{N}$ becomes excessively large. Upon analysis, it is evident that within a smaller range of $\mathcal{N}$, the fuzzy neighborhood primarily consists of neighboring samples that share highly similar semantics with the current text instance, thereby enhancing the consistency among samples within the same cluster. However, when $\mathcal{N}$ becomes too large, the fuzzy neighborhood may include instances from other clusters, introducing noise that disrupts the clustering objective. Based on empirical observations, a value of $\mathcal{N}$ in the range of 20 to 30 is recommended. Additionally, given that the Googlenews datasets are severely imbalanced with a large number of clusters, it is more sensitive to the choice of $\mathcal{N}$. Therefore, a



(a) ACC (b) NMI

Figure 4: Impact of hyperparameter $\mathcal{N}$ on the performance of the FNSCC model.

smaller value of $\mathcal{N}$ is set to 5.

5 Conclusion

This paper proposes a fuzzy neighborhood-aware self-supervised contrastive clustering (FNSCC) framework, comprising instance-level contrastive learning with neighborhood context and fuzzy neighborhood-aware clustering. It introduces two key innovations: leveraging neighborhood information to improve clustering confidence and reduce ambiguity, and removing false negatives within neighborhoods to enhance cluster-level discrimination, yielding a more suitable sample distribution for clustering.

Experimental results show that FNSCC outperforms other state-of-the-art models in terms of ACC and NMI on most short text datasets. Ablation studies further confirm the effectiveness of fuzzy neighborhood information in enhancing contrastive representation learning and clustering performance.

Limitations

The method proposed in this paper is similar to SCCL as it also requires the number of clusters to be set in advance. For datasets with many clusters and imbalanced distributions, FNSCC is more sensitive to the choice of neighbors. In future work, we plan to combine prior data knowledge with model characteristics to explore an adaptive mechanism for dynamically determining the optimal number of clusters and neighborhood range.

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

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

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.

Jinyu Cai, Shiping Wang, Chaoyang Xu, and Wenzhong Guo. 2022. Unsupervised deep clustering via contractive feature representation and focal loss. *Pattern Recognition*, 123:108386.

Billy Chiu, Sunil Kumar Sahu, Derek Thomas, Neha Sengupta, and Mohammady Mahdy. 2020. Autoencoding keyword correlation graph for document clustering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3974–3981.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724.

Benjamin Eckart, Wentao Yuan, Chao Liu, and Jan Kautz. 2021. Self-supervised learning on 3d point clouds by learning discrete generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8257.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. 2021. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pages 3015–3024. PMLR.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 55–65.

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.

Renchu Guan, Hao Zhang, Yanchun Liang, Fausto Giunchiglia, Lan Huang, and Xiaoyue Feng. 2020. Deep feature-based text clustering and its explanation. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3669–3680.

Vikram Gupta, Haoyue Shi, Kevin Gimpel, and Mrinmaya Sachan. 2022. Deep clustering of text representations for supervision-free probing of syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10720–10728.

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.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. 2022. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving Long Dialogue Summarization with Semantic Graph Representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883.

Hao Huang, Xiubo Geng, Pei Jian, Guodong Long, and Daxin Jiang. 2021. Reasoning over entity-action-location graph for procedural text understanding. In *ACL-IJCNLP 2021-59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

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.

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

Sheng-Chieh Lin and Jimmy Lin. 2023. A dense representation framework for lexical and semantic matching. *ACM Transactions on Information Systems*, 41(4):1–29.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

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.

N Reimers. 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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Sam Scott and Stan Matwin. 1998. Text classification using wordnet hypernyms. In *Usage of WordNet in natural language processing systems*.

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

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. pages 2840–2861.

Yepeng Sun, Jicang Lu, Ling Wang, Shunhang Li, and Ningbo Huang. 2023. Topic-Aware Contrastive Learning and K-Nearest Neighbor Mechanism for Stance Detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2362–2371.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, et al. 2024. Tnt-llm: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5836–5847.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782.

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.

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.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11556–11565.

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.

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

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

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.

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 61th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 10493–10507.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. 2021. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10875.

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 Additional Experiment Details

A.1 Datasets

The FNSCC algorithm is evaluated on six widely used benchmark text datasets. Table 2 presents detailed information about these datasets, including the number of documents, the number of categories, and the average word count per document.

Datasets	Size	Classes	Len(Avg)
AgNews	8000	4	23
SearchSnippets	12340	8	18
GoogleNews-TS	11109	152	28
GoogleNews-T	11109	152	6
GoogleNews-S	11109	152	22
Tweet	2472	89	8

Table 2: Statistics of text dataset. Size: number of documents; Classes: number of clusters; Len(Avg): Average number of words per document.

- **AgNews** (Zhang et al., 2015): A subset of 8000 news texts categorized into 4 different topics, collected and preprocessed by Rakib et al. (2020).
- **SearchSnippets** (Phan et al., 2008): Contains 12340 web search snippets covering 8 different domains.
- **GoogleNews** (Yin and Wang, 2016): Comprises 11109 news events across 152 categories, with the dataset segmented into full texts, titles, and text snippets, denoted as GoogleNews-TS, GoogleNews-T, and GoogleNews-S, respectively.
- **Tweet** (Yin and Wang, 2016): Consists of 2472 tweets distributed across 89 categories, sourced from the 2011-2012 microblog tracks of the text retrieval conference.

A.2 Baselines Description

The comparison baseline method used in this paper is described as follows:

- **BOW and TF-IDF** (Scott and Matwin, 1998; Bafna et al., 2016): Classical frequency-based methods that generate static text representations based on word occurrence statistics. BOW models texts as unordered word sets, while TF-IDF incorporates term specificity across documents to reduce the influence of common words.
- **STCC** (Xu et al., 2017): A representation learning method that integrates Word2Vec-based word embeddings with convolutional neural networks to capture local semantic patterns for short text clustering.

- **Self-train** (Hadifar et al., 2019): Autoencoders generate initial latent representations, which clustering supervision continuously refines to enhance feature quality for unsupervised tasks.

- **SBERT** (Reimers, 2019): An extension of BERT designed to produce semantically meaningful sentence embeddings using siamese and triplet networks, enabling more effective similarity measurement and clustering.

- **BGE-M3** (Xiao et al., 2024): A self-distillation strategy refines sentence embeddings under varied retrieval scenarios, improving robustness and generalization for clustering and matching.

- **SCCL** (Zhang et al., 2021): A contrastive clustering framework that jointly optimizes instance-level contrastive loss and a KL divergence loss over cluster assignment distributions to improve cluster compactness and separability.

- **ProPos** (Huang et al., 2022): Contrastive clustering is enhanced through prototype consistency and neighborhood alignment, yielding more discriminative and stable cluster representations.

- **CLSESSP** (Shen et al., 2024): The use of semantically enriched prototypes and divergence minimization strengthens sentence embedding discrimination in contrastive clustering.

- **Multi-MCCR** (Zhou et al., 2023): A semi-supervised method that leverages multi-model consistency and a contrastive BiKL divergence loss to refine pseudo-labels and enhance clustering accuracy.

- **RSTC** (Zheng et al., 2023): Adaptive optimal transport is integrated with contrastive learning under a pseudo-labeling regime to improve robustness against label noise and reinforce cluster structures.

A.3 Evaluation Metrics

This paper evaluates the clustering results of short texts using two widely adopted clustering performance metrics: clustering accuracy (ACC) and normalized mutual information (NMI). ACC is calculated using the Hungarian algorithm, which maximizes the matching between predicted and true labels. NMI quantifies the dependence between the clustering results and true labels by calculating the mutual information between them, normalized by their respective entropies. Both metrics have a value range of [0, 1], with higher values indicating better clustering performance.

ACC is defined as:

$$ACC = \frac{\sum_{i=1}^{N} \mathbb{1}_{(y_i = map(\hat{y}_i))}}{N}, \tag{10}$$

where $\mathbb{1}_{(y_i = map(\hat{y}_i))}$ is an indicator function, $map(\cdot)$ is the Hungarian mapping function, y_i and $\hat{y}_i$ represent the real label and predicted label, respectively. NMI is defined as:

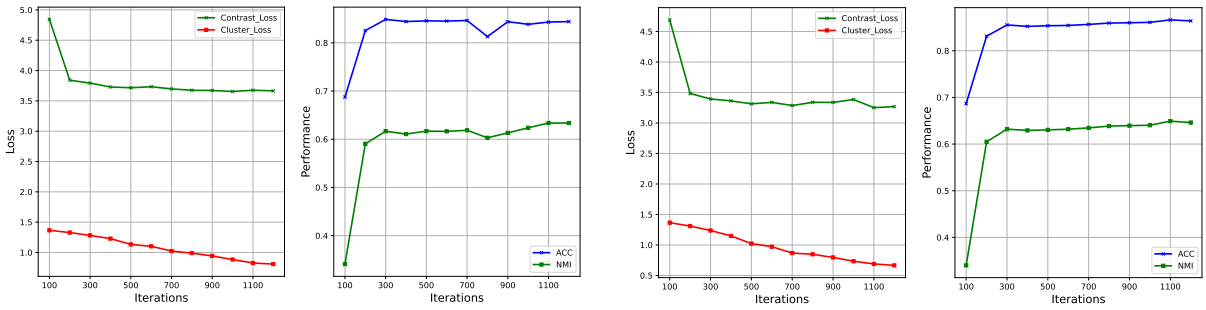
$$NMI = \frac{2 \cdot MI(Y, \hat{Y})}{H(Y) + H(\hat{Y})}, \tag{11}$$

where Y and $\hat{Y}$ represent the distribution sequences of real labels and predicted labels, $MI(Y, \hat{Y})$ represents the mutual information between Y and $\hat{Y}$, and $H(\cdot)$ represents information entropy.

A.4 Hyperparameters Description

- $\boldsymbol{\alpha}$**&**$\boldsymbol{\beta}$ **:** The parameters α and β are the weights used to balance the instance contrastive loss and clustering loss, respectively. In the total loss, the instance contrastive loss primarily ensures effective separation between the embedded representations of different samples, thereby enhancing the model's discriminative capability. On the other hand, the clustering loss focuses on enabling the model to learn the clustering structure more effectively by grouping similar samples together, improving both the accuracy and consistency of clustering results. By fine-tuning α and β, an optimal trade-off can be achieved between instance discrimination and clustering performance. In this paper, experimental tuning is performed, with the default values of α and β set to 10 and 1 (1 and 1 for the Tweet dataset), respectively.

- $\boldsymbol{\tau}$ **:** The temperature coefficient τ is used to scale the numerical range of similarity between samples, thereby adjusting the model's sensitivity to differences in similarity. A smaller τ amplifies the impact of similarity,



(a) SCCL (b) FNSCC

Figure 5: Training loss and clustering performance of SCCL and FNSCC on the AgNews dataset.

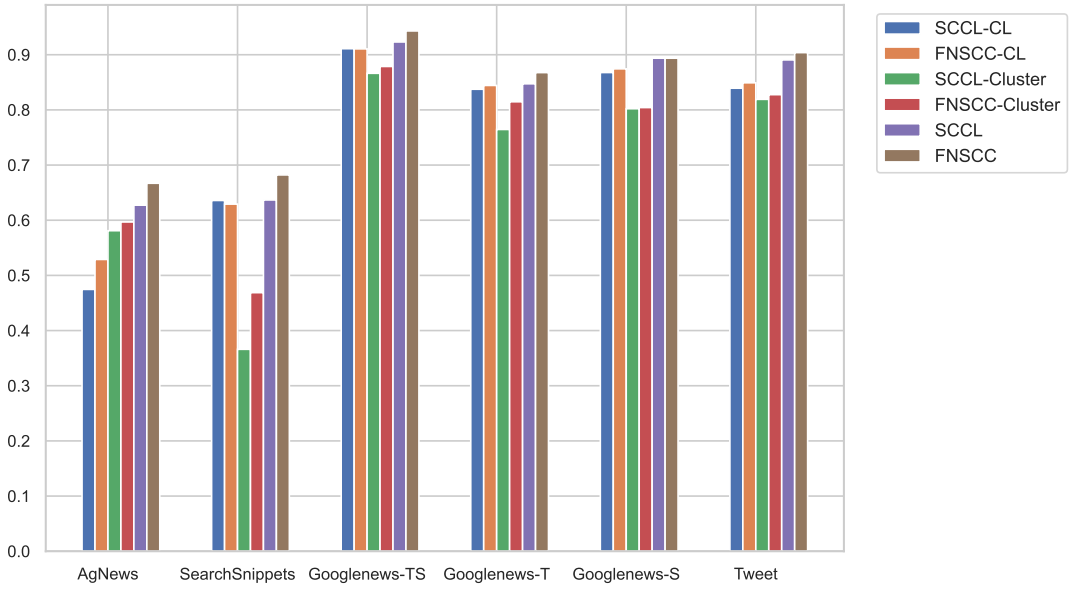


Figure 6: Comparison of ablation studies on NMI between FNSCC and SCCL.

causing the model to emphasize sample pairs with high similarity. Conversely, a larger τ smooths the impact of similarity, reducing the model's sensitivity to differences among sample pairs. In this paper, τ is set to a default value of 0.5.

- $\boldsymbol{\gamma}$ **:** γ represents the degrees of freedom of the Student's t-distribution. Following the SCCL framework, γ is fixed at 1 in this work.

- $\boldsymbol{\lambda}$ **:** λ is an neighbor weight coefficient that controls the impact of neighbor information in updating q_{ik}. A larger λ increases the importance of neighbor information, promoting greater consideration of local consistency during clustering. Conversely, a smaller λ reduces the impact of neighbor information, emphasizing the characteristics of the instance itself. In this paper, λ is set to 0.5 by default.

- $\boldsymbol{\mathcal{N}}$ **:** The size of the neighborhood $\mathcal{N}$ for a text instance significantly affects the fuzzy neighborhood-aware mechanism. Larger $\mathcal{N}$ values provide more neighbor information, potentially improving local consistency, but may also introduce noise if irrelevant neighbors are included. A detailed analysis of $\mathcal{N}$ and its impact on clustering performance is provided in the experimental section. Furthermore, this paper sets the number of neighbors for both the clustering and instance comparison heads to ensure consistency in their modeling of the local data structure. Specifically, the negative samples excluded in the instance comparison head are treated as similar objects within the neighborhood and are incorporated into the clustering head calculation, facilitating information sharing and semantic alignment between the two task heads.

	AgNews		SearchSnippets		GoogleNews-TS	
Substitution rate	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
(BERT) %10	84.29	65.32	78.32	67.13	86.13	93.29
(BERT) %20	86.43	65.82	80.63	67.32	86.82	94.21
(BERT) %30	85.96	65.48	79.01	66.82	85.21	93.24
(RoBERTa) %10	85.32	65.88	81.65	67.17	85.23	93.18
(RoBERTa) %20	86.13	65.92	82.13	66.94	87.77	93.64
(RoBERTa) %30	84.97	64.13	78.63	65.93	85.58	92.97
(BERT+RoBERTa) %10	85.56	65.48	79.01	66.29	86.42	94.10
(BERT+RoBERTa) %20	**87.85**	**66.70**	**82.59**	**67.65**	**88.21**	**94.31**
(BERT+RoBERTa) %30	86.91	65.63	79.21	66.87	85.27	93.83
	GoogleNews-T		**GoogleNews-S**		**Tweet**	
Substitution rate	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
(BERT) %10	68.83	85.32	79.64	88.97	80.94	89.54
(BERT) %20	71.19	86.22	79.54	88.68	81.75	89.68
(BERT) %30	70.85	85.76	78.96	87.82	80.45	88.84
(RoBERTa) %10	69.73	86.11	78.24	87.13	78.92	88.26
(RoBERTa) %20	71.72	86.21	78.97	88.14	82.74	89.57
(RoBERTa) %30	70.27	86.02	77.18	87.07	79.31	89.14
(BERT+RoBERTa) %10	70.79	86.09	80.09	88.67	81.23	89.74
(BERT+RoBERTa) %20	**72.72**	**86.76**	**80.49**	**89.37**	**83.62**	**90.38**
(BERT+RoBERTa) %30	69.05	85.00	79.33	89.24	81.07	89.63

Table 3: Comparison of FNSCC Clustering Performance on Real Text Datasets under Different Contextual Augmenter Settings.

A.5 Training Loss and Clustering Performance

To provide a more comprehensive comparison, we present the changes in loss values and model performance on the AgNews dataset over training iterations, as shown in Figure 5. This analysis is based on visual results from the training process.

The results in Figure 5 show that the loss and performance trends for both FNSCC and SCCL are similar during training. However, FNSCC outperforms SCCL, achieving better values and demonstrating greater stability after convergence. This highlights that integrating fuzzy neighborhood information into the SCCL training process allows the model to better align text embeddings with clustering objectives and more effectively separate negative samples.

A.6 The Impact of Each Component in FNSCC on the NMI Index.

Figure 6 presents the NMI, a clustering metric, for six short text datasets under different components of SCCL and FNSCC. The results align with the ACC trend in Section 4.3, showing that fuzzy neighborhoods significantly improve the NMI of both the instance contrastive head and the clustering head across most datasets.

A.7 Discussion on Data Augmentation

Data augmentation plays a crucial role in the contrastive learning module by generating different types of contrasting instances, significantly impacting model performance. Zhang et al. (2021) systematically investigates various unsupervised text augmentation methods, demonstrating through extensive experiments the superior efficacy of a *Contextual Augmenter*. This *Augmenter* leverages a pre-trained Transformer model to identify the top-n suitable words in the input text for substitution. Given that FNSCC also adopts a Transformer-based encoder, we employ the same augmentation strategy to ensure compatibility and semantic consistency within the framework.

Furthermore, this study uses two distinct masked language models for context-based substitution augmentation. To investigate the impact of different MLMs and substitution rate on model performance, experiments are conducted on the datasets listed in Table 2, comparing performance across

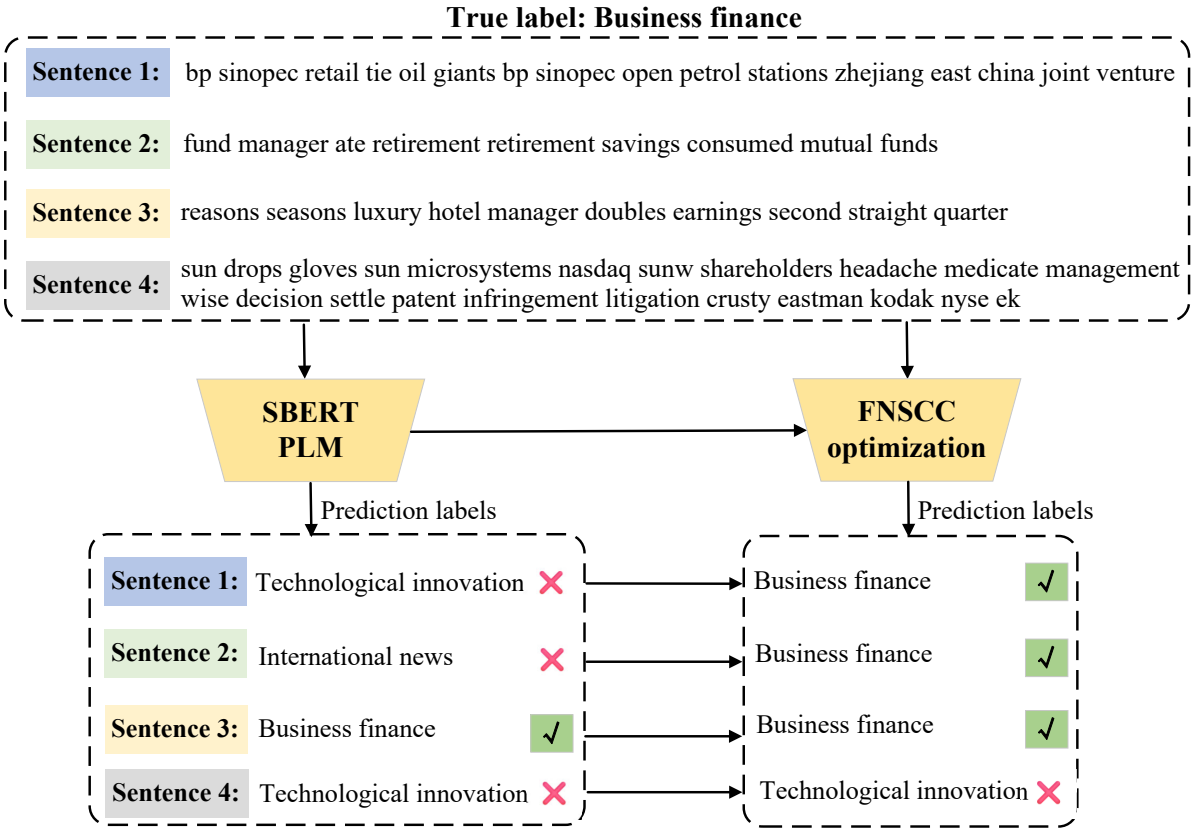


Figure 7: Case studies of short texts in the AgNews dataset. The true label for all four sentences is "business finance." The clustering results based on the pre-trained SBERT model only correctly predict sentence 3. After optimization with FNSCC, most sentences are correctly clustered. Nevertheless, due to the ambiguity of cluster boundaries, a small number of incorrect predictions still exist.

various substitution configurations. The results, summarized in Table 3, show that FNSCC achieves optimal performance at a substitution rate of 20% when using a mixture of BERT and RoBERTa.

This finding provides an empirical guideline, showing that applying different MLMs with a 20% substitution rate achieves an optimal balance between semantic enrichment and distortion. This approach effectively broadens the data distribution while minimizing the introduction of excessive noise or semantic drift from the original text.

A.8 Analysis of Failure Cases

Despite its effectiveness in separating clusters and grouping similar instances, as shown in Figure 2 (c), FNSCC still encounters some errors near cluster boundaries. We analyze these failure cases from both algorithmic and empirical perspectives, and summarize them as follows:

(1) Semantic ambiguity at cluster boundaries

The data categories in AgNews mainly include "business finance", "technological innovation", "sports news", and "international news". Among them, there are many samples of similar categories that are often confused. For example, short texts of "business finance" and "technological innovation" have similar vocabulary and structure, which leads to confusion in the embedding space. As shown in Figure 7, using samples from the AgNews dataset, the true label for all four short texts is "business finance". FNSCC corrects most of the misclassifications by introducing instance-level and cluster-level fuzzy neighborhood optimization on the SBERT pre-trained model. However, the error in sentence 4 remains uncorrected. Analysis of the adjacency structure of these samples shows strong semantic overlap between neighborhoods, suggesting that more refined discrimination strategies are still needed in low-density regions. Future research explores methods based on density or local uncertainty to address this issue.

(2) Class imbalance and neighborhood domination

When a category such as "business finance" has a high proportion, it will dominate the neighborhood of nearby instances, even including minority

categories such as “technological innovation” or “international news”. However, due to the large number and similar terms, its surroundings are easily drowned out by business finance samples. For example, when the classes are unbalanced, there will be samples in the neighborhood that do not match the current class. In Equation (5), $\sum_{j \in \mathcal{N}(g(t_i))} \Psi(t_i, t_j) \cdot q_{jk}$ will be affected by samples from the wrong category to distort the fuzzy neighborhood assignment, which further affects the target distribution (Equation (7)). Although the adaptive weight $\Psi(t_i, t_j) = s(g(t_i), g(t_j)) \cdot d(g(t_i), g(t_j))$ will regulate it overall, it will still be affected to a certain extent.

A.9 Computational Cost

The training process for each dataset is conducted on a GeForce RTX 4090 GPU, with an approximate runtime ranging from 10-30 minutes.