ISCST: Improved Supporting Clustering Based on Sentence-Transformers with Contrastive Learning

Kongqiang Wang, Xuejie Zhang, Jin Wang, Xiaobing Zhou*
School of Information Science and Engineering, Yunnan University, Kunming, China
* Corresponding author: zhouxb@ynu.edu.cn.

***Abstract*—Unsupervised clustering faces challenges when semantic categories overlap in the representation space, limiting the effectiveness of distance-based methods. We propose an Improved Supporting Clustering based on Sentence-Transformers with Contrastive Learning (ISCST), a novel framework combining bottom-up instance discrimination with top-down clustering objectives. ISCST leverages contrastive learning to simultaneously enhance categorical separation and optimize intra-cluster compactness and inter-cluster discrimination. Evaluated on multiple short text clustering benchmarks, ISCST achieves substantial improvements over state-of-the-art methods: 5%-8% Accuracy enhancement and 3%-5% Normalized Mutual Information improvement. A comprehensive quantitative analysis validates ISCST's effectiveness in addressing category overlap, producing semantically meaningful and well-separated clusters in the learned representation space. Our joint optimization approach synergizes contrastive learning with clustering objectives to achieve superior performance. The source code is available at https://github.com/WangKongQiang/ISCST.**

***Index Terms*—Sentence Transformers, Contrastive Learning, Unsupervised Clustering, Short Text Clustering**

I. Introduction

Clustering represents one of the most fundamental challenges in unsupervised machine learning, attracting extensive research attention over decades. Traditional algorithms like K-means [1] and Gaussian Mixture Models [2] measure distances in the original data space, but prove inadequate for high-dimensional data, where the curse of dimensionality significantly degrades performance. On the other hand, deep neural networks have opened new avenues by enabling the transformation of raw data into lower-dimensional, discriminative representation spaces. Recent research increasingly integrates clustering objectives with deep representation learning, performing clustering operations by optimizing a clustering objective defined in learned feature spaces [3]. Despite promising improvements, the clustering performance is still inadequate, especially in the presence of complex data with a large number of clusters. However, clustering performance remains suboptimal for complex datasets, as sophisticated neural networks often fail to eliminate substantial category overlap in representation space before clustering initialization.

On the other hand, Instance-wise Contrastive Learning (Instance-CL) [11] has recently achieved remarkable success in self-supervised learning. Instance-CL usually optimizes on an auxiliary set obtained by data augmentation. As the name suggests, a contrastive loss is then adopted to pull together samples augmented from the same instance in the original dataset while pushing apart those from different ones. Essentially, Instance-CL disperses different instances apart while implicitly bringing similar instances together to some extent. This beneficial property can be leveraged to support clustering by separating overlapping categories. Then clustering, thereby better separates different clusters while tightening each cluster by explicitly bringing samples in that cluster together. As illustrated in Figure 1, one possible reason is that, even with a deep neural network, data still has significant overlap across categories before clustering starts. Consequently, the clusters learned by optimizing various distance or similarity-based clustering objectives suffer from poor purity. Although the most advanced research methods, such as SCCL and DACL, could perform clustering separation to a certain extent, they still could not clearly and distinctly separate clusters with distinct boundaries.

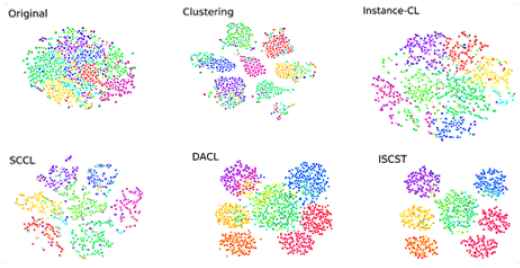

Fig. 1. TSNE visualization of the embedding space learned on SearchSnippets using Sentence Transformer Sentence-bert [17] as backbone. Each color corresponds to a ground-truth semantic category.

To this end, we propose Improved Supporting Clustering Based on Sentence-Transformers with Contrastive Learning (ISCST), which jointly optimizes a top-down clustering loss and a bottom-up instance-wise contrastive loss. We assess the performance of ISCST for short-text clustering, which has become increasingly important given the popularity of social media such as Twitter and Instagram. It benefits many real-world applications, including topic discovery [12], recommendation [13], and visualization [14].

II. Related Work

A. Self-supervised learning

Self-supervised learning has emerged as a dominant paradigm for learning effective representations without manual annotation. Early approaches focused on pretext tasks like

predicting masked tokens [4], generating future tokens [5], or denoising corrupted tokens [6] for text, and colorization [7], rotation [8], or patch positioning [9] for images. However, these methods produced task-specific representations with limited generalization. Instance-wise Contrastive Learning (Instance-CL) [11] has emerged as a breakthrough paradigm in self-supervised representation learning. Instance-CL constructs augmented sample pairs and employs contrastive loss functions to maximize similarity between augmented versions while minimizing similarity between different instances, effectively dispersing distinct instances while encouraging aggregation of semantically similar instances. Instance-wise contrastive learning (Instance-CL) has driven recent breakthroughs. Inspired by Becker and Hinton [15], Instance-CL treats each instance and its augmentations as independent classes, pulling together augmented versions while pushing apart different instances [16]. This produces well-separated embeddings while preserving local invariances.

Despite success, Instance-CL has limitations. While implicitly grouping similar instances [11], it indiscriminately separates different instances regardless of semantic similarity. This instance-centric objective creates unstable, data-dependent grouping, potentially leading to suboptimal representations [18]. This particularly hinders clustering, where discovering semantic categories matters more than distinguishing individual instances.

B. Short Text Clustering

Short text clustering presents unique challenges due to limited semantic information per instance. In this scenario, traditional BoW and TF-IDF approaches produce sparse vectors that lack expressive ability. To remedy this issue, early neural network solutions were proposed to enrich representations [19], incorporating word embeddings to further enhance performance. [20], However, the above approaches suffered from multi-stage optimization requirements with independent component training.

On the other hand, despite the tremendous successes of contextualized word embeddings [21] in NLP tasks, their potential for short-text clustering remains unexplored. We bridge this gap by using pre-trained sentence transformers as a backbone, enabling end-to-end optimization.

Building upon these insights, we propose an Improved Supporting Clustering based on Sentence-Transformers with Contrastive Learning (ISCST), synergistically combining top-down clustering objectives with bottom-up instance-wise contrastive learning. We evaluate our method for short text clustering, an increasingly important task for social media applications, including automated topic discovery, personalized recommendation, and data visualization.

III. Method

We propose a joint model that combines instance-wise contrastive learning (Instance-CL) and clustering for unsupervised representation learning. As shown in Figure 2, a backbone $\psi(\cdot)$ maps the input data to embeddings, followed by two different heads: $g(\cdot)$ for contrastive loss and $f(\cdot)$ for clustering loss. Our data consists of both the original and the augmented data. Specifically, for a randomly sampled minibatch $\mathcal{B} = \{x_i\}_{i=1}^{M}$, we randomly generate augmented data $\mathcal{B}^a = \{\tilde{x}_i\}_{i=1}^{2M}$.

A. Instance-wise Contrastive Learning

For $\tilde{x}_{i^1}, \tilde{x}_{i^2} \in \mathcal{B}^a$ from the same instance (positive pair), the loss is

$$\ell_{i^1}^{I} = -\log \frac{\exp(\text{sim}(\tilde{z}_{i^1}, \tilde{z}_{i^2})/\tau)}{\sum_{j=1}^{2M} \mathbf{1}_{j \neq i^1} \exp(\text{sim}(\tilde{z}_{i^1}, \tilde{z}_j)/\tau)}, \tag{1}$$

with $\tilde{z}_j = g(\psi(\tilde{x}_j))$, $\tau = 0.5$, and

$$\text{sim}(\tilde{z}_i, \tilde{z}_j) = \frac{\tilde{z}_i^\top \tilde{z}_j}{\|\tilde{z}_i\|_2 \|\tilde{z}_j\|_2}. \tag{2}$$

The overall contrastive loss is

$$\mathcal{L}_{\text{Instance-CL}} = -\frac{1}{2M} \sum_{i=1}^{2M} \log \frac{\exp(\text{sim}(\tilde{z}_{i^1}, \tilde{z}_{i^2})/\tau)}{\sum_{j=1}^{2M} \mathbf{1}_{j \neq i^1} \exp(\text{sim}(\tilde{z}_{i^1}, \tilde{z}_j)/\tau)}. \tag{3}$$

B. Clustering

We simultaneously encode the semantic categorical structure into the representations via unsupervised clustering. Unlike Instance-CL, clustering focuses on the high-level semantic concepts and aims to group instances from the same semantic category. Let $e_j = \psi(x_j)$ be the embedding of x_j and μ_k the centroid of cluster k. The soft assignment is computed with the Student's t-distribution [22]:

$$q_{jk} = \frac{(1+\|e_j - \mu_k\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K} (1+\|e_j - \mu_{k'}\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}}, \tag{4}$$

where $\alpha = 1$. Following Junyuan Xie [3], the auxiliary target distribution is

$$p_{jk} = \frac{q_{jk}^2 / f_k}{\sum_{k'} q_{jk}^2 / f_{k'}}, \quad f_k = \sum_{j=1}^{M} q_{jk}. \tag{5}$$

The clustering loss is defined as

$$\mathcal{L}_{\text{Cluster}} = \frac{1}{M} \sum_{j=1}^{M} \sum_{k=1}^{K} p_{jk} \log \frac{p_{jk}}{q_{jk}}. \tag{6}$$

C. Overall Objective

Our final loss combines the two components:

$$\mathcal{L} = \mathcal{L}_{\text{Instance-CL}} + \eta \, \mathcal{L}_{\text{Cluster}}, \tag{7}$$

with $\eta = 10$ balancing contrastive and clustering objectives.

IV. Experiments

We implement our model in PyTorch [23] with the Sentence Transformer library [17]. ISCST employs all-mpnet-base-v2 for embedding performance. Both utilize linear clustering heads (f) of dimensions $768 \times K$, where K represents the number of target clusters. The contrastive learning component employs an MLP (g) with 768 hidden units and 128-dimensional outputs.

Following established protocols [24], we evaluate using the Accuracy and NMI metrics. Results demonstrate significant improvements: 5%-8% in Accuracy and 3%-5% in NMI over state-of-the-art methods across eight benchmark datasets, as detailed in Table 1.

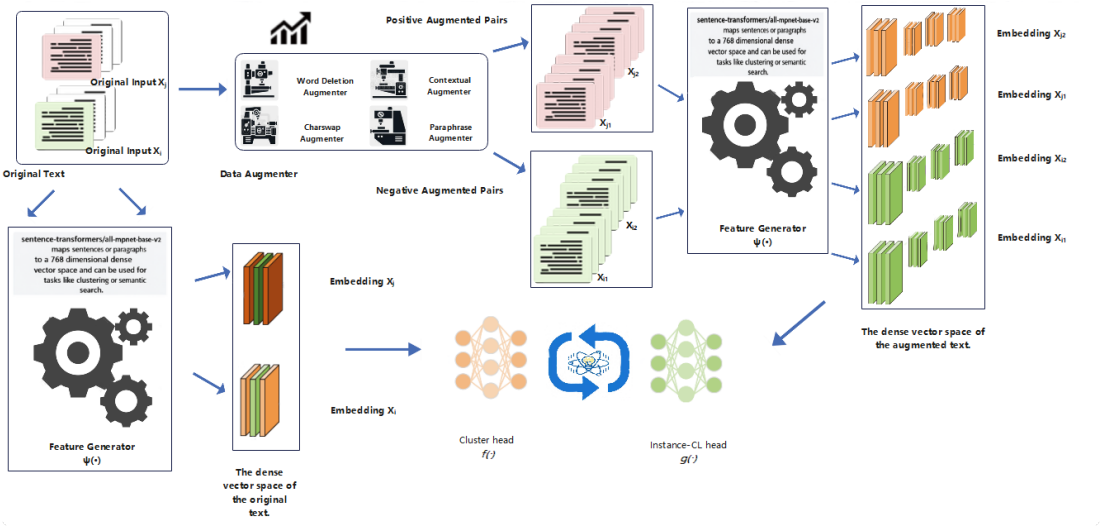

Fig. 2. Training framework of ISCST. We jointly optimize the clustering loss on the original data and the contrastive loss on augmented pairs.

TABLE I
EXISTING SHORT TEXT DATASET STATISTICS

Dataset	$\|V\|$	Documents		Clusters	
		N^D	Len	N^C	L/S
AgNews	21K	8000	23	4	1
StackOverflow	15K	20000	8	20	1
Biomedical	19K	20000	13	20	1
SearchSnippets	31K	12340	18	8	7
GoogleNewsTS	20K	11109	28	152	143
GoogleNewsS	18K	11109	22	152	143
GoogleNewsT	8K	11109	6	152	143
Tweet	5K	2472	8	89	249

$|V|$: vocabulary size; N^D: number of short text documents; Len: average number of words per document; N^C: number of clusters; L/S: ratio of the largest cluster size to the smallest cluster size.

A. Exploration of Data Augmentations

We explore four unsupervised text augmentation strategies: (1) *Word Deletion Augmenter*. It randomly removes words using WordNet [25]; (2) *Charswap Augmenter*. It enhances text by swapping random characters of words in a sentence; (3) *Contextual Augmenter* [26]. It leverages pre-trained transformers to identify top-n suitable words for insertion or substitution. For contextual augmentation, we use google-bert/bert-base-uncased and FacebookAI/roberta-base models to generate augmented pairs, providing transformer-based semantic understanding for robust positive pair creation in contrastive learning. (4) *Paraphrase Augmenter*. Machine translation models can be used to paraphrase text by translating it to an intermediate language and back (round-trip translation). This augmenter shows how to paraphrase text by first passing it through an English-French translation model, then through a French-English mixture-of-experts translation model. Using the example input "The new date for the Games, postponed for a year in response to the coronavirus pandemic, gives athletes time to recalibrate their training schedules" as basic text to generate two sentences obtained through paraphrasing. Data flow is shown in Figure 3.

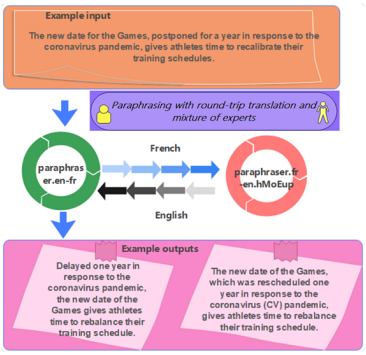

Fig. 3. Paraphrasing with round-trip translation and mixture of experts.

B. Comparison with State-of-the-art Methods

We compare against several established approaches for short text clustering evaluation.

STCC [24] employs three-stage optimization: Word2Vec pre-training on in-domain corpus, CNN-based representation enrichment, and K-means clustering. **Self-Train** [19] enhances pretrained embeddings using layer-wise pretrained autoencoders with clustering objectives and interval-based target distribution updates. **SimCTC** [27] proposes simple contrastive learning for text clustering improvement. **HAC-SD** [28] applies hierarchical agglomerative clustering on sparse pairwise similarity matrices with threshold-based pruning. **BoW&TF-IDF** applies K-means on 1500-dimensional traditional features. **SCCL** [29] introduces the Supporting Clustering with Contrastive Learning framework for better separation. **DACL** [30] improves SCCL through pseudo-label enhancement.

Our method achieves superior performance by updating target distributions at each iteration, demonstrating significant improvements over multi-stage and hierarchical approaches.

C. Dataset Statistic

We evaluate ISCST on eight benchmark datasets spanning diverse domains (see Table 1). **SearchSnippets** [31] contains 12,340 web search snippets across 8 groups. **StackOverflow** includes 20,000 technical question titles in 20 categories, extracted from Kaggle [24]. **Biomedical** comprises 20,000 scientific paper titles from 20 research domains, randomly selected from the PubMed corpus distributed by BioASQ [24]. **AgNews** covers news titles in 4 topics [28]. **Tweet** contains 2,472 social media posts across 89 categories [32]. GoogleNews includes 11,109 articles across 152 events, partitioned into **GoogleNews-TS** (full), **GoogleNews-T** (titles), and **GoogleNews-S** (snippets) variants [28].

For contrastive learning, we employ the Contextual Augmenter [26] with optional character-swap, the WordNet Augmenter with word deletion and character swap, or the Paraphrase Augmenter, selected based on dataset characteristics to optimize positive pair generation.

D. Evaluation Metrics

We employ four standard clustering metrics: Accuracy (**ACC**) measures agreement between predicted assignments and ground truth categories. Normalized Mutual Information (**NMI**) quantifies information correlation between clustering results and true labels (0-1 range). Adjusted Rand Index (**ARI**) evaluates consistency while correcting for chance (-1 to 1 range). Adjusted Mutual Information (**AMI**) functions like NMI with random clustering adjustment (0-1 range). Based on the short text clustering complexity, we prioritize ACC and NMI for model comparison, providing complementary perspectives through direct category matching and information-theoretic relationships, respectively.

V. Results and Analysis

A. Approach Comparison

We demonstrate robustness by avoiding preprocessing on all datasets, while most baselines employ extensive text normalization procedures.

Table 2 shows that ISCST achieves superior performance across most datasets. Notable exceptions provide valuable insights: Hadifar et al. [19] outperform on Biomedical due to domain-specific word embeddings trained on biomedical corpora and layer-wise pretrained autoencoders, contrasting with our general-purpose sentence transformers.

Importantly, our approach offers superior scalability for large-scale applications.

B. Augmentation Results Comparison

We evaluate augmentation strategies using WordNet Augmenter (word deletion, character-swap), Contextual Augmenter (transformer-based word insertion/substitution) and Paraphrase Augmenter. For both WordNet Augmenter and Contextual Augmenter, we test three different settings by choosing the word substitution ratio for each text instance to 10%, 20%, and 30%, respectively. After extensive experiments across multiple datasets, it was found that a 30% word substitution ratio per text instance was ineffective because it significantly altered the original meaning of the instance. Only in longer text instances does it show a small advantage, such as in GoogleNews-TS, because each text instance has 28 words (see Table 1). So that the data instances, after undergoing the augmentation strategies, can still maintain their semantics very well. For most datasets, combining methods with 10% and 20% occurrence proportions balances diversity with semantic preservation. Table 3 shows that these proportional-combinations augmentation strategies and Paraphrase Augmenter yield significant improvements in ISCST performance. The results have verified that our method for clustering short texts has significant advantages. Figure 4 (*Right*) shows the impact of using a composition of data augmentations, including the Contextual Augmenter and CharSwap Augmenter. As we can see, using a composition of data augmentations does boost the performance of ISCST on GoogleNews-TS, where the average number of words in each text instance is 28. However, we observe the opposite on StackOverflow, where the average number of words in each instance is 8. This result differs from what has been observed in the image domain, where using a composition of data augmentations is crucial for contrastive learning to attain good performance. A possible explanation is that generating high-quality augmentations for textual data is more challenging, since changing a single word can invert the semantic meaning of the entire instance. This challenge is compounded when a second round of augmentation is applied to very short text instances, such as StackOverflow. We further demonstrate this in Figure 4 (*Left*), where the augmented pairs of StackOverflow largely diverge from the original texts in the representation space after the second round of augmentation.

TABLE II
Clustering results on eight short text datasets

Method	AgNews		SearchSnippets		StackOverflow		Biomedical	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	27.6	2.6	24.3	9.3	18.5	14.0	14.3	9.2
TF-IDF	34.5	11.9	31.5	19.2	58.4	58.7	28.3	23.2
STCC	-	-	77.0	63.2	51.1	49.0	43.6	38.1
Self-Train	-	-	77.1	56.7	59.8	54.8	**54.8**	**47.1**
SimCTC	-	-	85.4	71.1	78.3	75.4	-	-
HAC-SD	81.8	54.6	82.7	63.8	64.8	59.5	40.1	33.5
SCCL	88.2	68.2	85.2	71.1	75.5	74.5	46.2	41.5
DACL	88.6	69.0	86.1	73.6	77.5	76.0	48.6	40.3
ISCST(Ours)	**90.1**	**71.2**	**88.3**	**75.2**	**84.7**	**78.6**	52.3	45.2
Method	GoogleNews-TS		GoogleNews-T		GoogleNews-S		Tweet	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	57.5	81.9	49.8	73.2	49.0	73.5	49.7	73.6
TF-IDF	68.0	88.9	58.9	79.3	61.9	83.0	57.0	80.7
STCC	-	-	-	-	-	-	-	-
Self-Train	-	-	-	-	-	-	-	-
SimCTC	90.9	96.1	-	-	-	-	84.7	93.4
HAC-SD	85.8	88.0	81.8	84.2	80.6	83.5	89.6	85.2
SCCL	89.8	94.9	75.8	88.3	83.1	90.4	78.2	89.2
DACL	89.6	94.4	80.5	89.8	84.0	91.7	82.0	90.6
ISCST(Ours)	**91.1**	**96.3**	**85.4**	**93.7**	**85.3**	**93.0**	**92.2**	**95.6**

Results are averaged over ten random runs.

C. Ablation Study

We conduct systematic ablation experiments analyzing individual components and optimization strategies. We evaluate three variants: standalone clustering component (Clustering), instance-wise contrastive learning alone (Instance-CL), and sequential training (ISCST-Seq) with Instance-CL pre-training followed by clustering optimization.

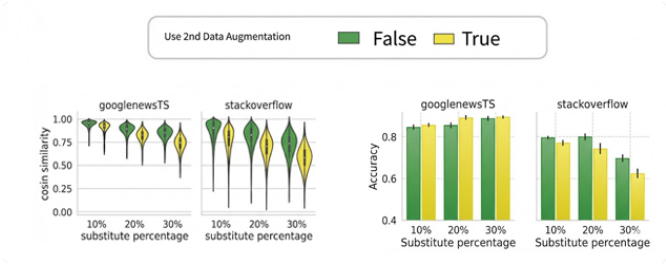

Fig. 4. Impact of using composition of data augmentations. Either only Contextual Augmenter (green) is used, or Contextual Augmenter and CharSwap Augmenter are applied sequentially (yellow). (*Right*) Clustering accuracy versus variant augmentation strengths, the x-axis indicates the percentage of words in each instance being changed by the associated data augmentation technique. (*Left*) Distribution of the cosine similarity between the representations of each original text and its augmented pair at the beginning of training.

TABLE III
CLUSTERING RESULTS ON EIGHT SHORT TEXT DATASETS

Data Augmentation Method	AgNews				SearchSnippets				StackOverflow				Biomedical			
	ACC	NMI	ARI	AMI	ACC	NMI	ARI	AMI	ACC	NMI	ARI	AMI	ACC	NMI	ARI	AMI
charswap_10	88.6	69.3	72.9	69.3	86.2	72.4	72.1	72.3	**84.7**	**78.6**	**68.6**	**78.6**	50.8	43.9	30.9	43.8
charswap_20	88.7	69.6	73.1	69.5	78.9	67.4	63.7	67.3	82.9	77.9	60.7	77.8	50.6	43.7	30.8	43.6
word_deletion_10	88.8	69.8	73.3	69.8	79.3	68.1	64.8	68.0	73.1	66.9	58.9	66.8	50.3	43.5	29.9	43.3
word_deletion_20	88.9	70.0	73.4	69.9	79.6	68.4	65.2	68.4	72.9	66.7	58.6	66.6	50.3	42.6	29.5	42.4
trans_subst_10	88.3	68.9	72.2	68.9	85.7	71.2	70.9	71.1	79.3	74.6	65.8	74.6	49.2	41.3	29.0	41.1
trans_subst_20	88.5	69.2	72.6	69.2	86.4	73.9	74.5	73.1	80.2	78.0	68.3	77.9	49.1	41.2	28.9	41.0
trans_subst_10_charswap_10	88.2	68.9	72.3	68.9	85.9	71.6	71.8	71.6	77.2	71.9	63.2	71.8	49.2	41.3	29.0	41.1
trans_subst_20_charswap_20	88.5	68.9	72.7	68.9	85.2	71.0	72.3	71.0	74.5	70.3	60.3	70.2	49.2	41.3	29.0	41.1
paraphrase	**90.1**	**71.2**	**74.6**	**71.1**	**88.3**	**75.2**	**76.6**	**75.2**	82.8	77.7	60.6	77.5	**52.3**	**45.2**	**32.1**	**45.2**
Data Augmentation Method	**GoogleNews-TS**				**GoogleNews-T**				**GoogleNews-S**				**Tweet**			
	ACC	NMI	ARI	AMI	ACC	NMI	ARI	AMI	ACC	NMI	ARI	AMI	ACC	NMI	ARI	AMI
charswap_10	84.6	93.7	80.8	92.6	76.6	88.7	72.4	86.6	79.1	89.9	75.0	88.1	83.9	91.0	88.3	88.9
charswap_20	85.4	93.6	82.1	92.5	**85.4**	**93.7**	**82.1**	**92.6**	78.3	89.4	75.8	87.5	85.2	91.9	88.6	89.9
word_deletion_10	85.5	93.8	82.7	92.6	78.1	89.5	79.2	87.8	80.9	90.4	76.0	88.7	79.5	91.5	81.5	89.2
word_deletion_20	87.5	94.4	83.8	93.3	79.5	89.7	76.5	87.9	82.2	90.7	78.7	89.0	79.5	91.6	76.8	89.3
trans_subst_10	85.6	94.2	83.6	93.1	80.6	90.2	78.9	88.4	81.8	90.7	77.3	89.0	75.9	90.1	73.7	87.6
trans_subst_20	86.7	94.4	83.0	93.5	81.4	90.4	80.5	88.7	**85.3**	**93.0**	**80.8**	**91.3**	85.9	93.3	86.2	91.6
trans_subst_10_charswap_10	87.0	94.3	84.4	93.3	78.0	89.0	76.4	87.0	82.1	90.7	78.6	89.1	86.4	92.9	91.9	91.1
trans_subst_20_charswap_20	**91.1**	**96.3**	**88.1**	**94.7**	79.6	89.4	79.4	87.5	78.3	89.4	75.7	87.5	90.1	94.5	95.4	93.0
paraphrase	89.2	94.2	86.3	92.5	83.2	91.5	80.1	90.8	83.2	91.2	78.6	89.1	**92.2**	**95.6**	**96.1**	**94.8**

All reported results are averaged over ten random runs.

Figure 5 shows that Instance-CL can group semantically similar instances, but this effect remains implicit and dataset-dependent, leading to inconsistent performance across domains and cluster distributions.

Our joint optimization approach (ISCST) consistently outperforms both individual components by substantial margins, demonstrating synergistic benefits of combining bottom-up instance discrimination with top-down clustering objectives. Notably, ISCST outperforms its sequential counterpart (ISCST-Seq), underscoring the critical importance of simultaneous optimization.

Results validate our hypothesis that joint optimization enables mutual reinforcement: contrastive learning scatters overlapping categories while clustering objectives encourage intra-cluster cohesion and inter-cluster separation. Joint training's superiority confirms that simultaneous optimization is essential for leveraging both paradigms' full potential.

To further investigate what enables better ISCST performance, we track both intra- and inter-cluster distances in the representation space throughout the learning process. For a given cluster, the intra-cluster distance is the average distance between the centroid and all samples grouped into that cluster, and the inter-cluster distance is the distance to its closest neighbor cluster. In Figure 6, we report each distance type and its mean, averaged over all clusters, where the clusters are defined either with respect to ground-truth labels (solid lines) or model-predicted labels (dashed lines). Figure 6 shows that clustering achieves a smaller intra-cluster distance and a

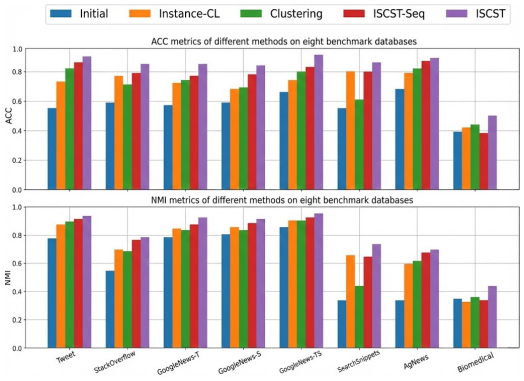

Fig. 5. Ablation study of ISCST. In ISCST-Seq, we first train the model using Instance-CL, and then optimize the clustering objective. We can see that the ISCST method performs best in the ACC and NMI metrics on the eight benchmark databases.

larger inter-cluster distance when evaluated on the predicted clusters. It demonstrates the ability of clustering to group each self-learned cluster and separate different clusters. However, we observe the opposite when evaluated on the ground-truth clusters, along with low Accuracy and NMI scores. One possible explanation is that data from different ground-truth clusters often overlap significantly in the embedding space before clustering begins (see the upper-left subgraph in Figure 1), making it hard for our distance-based clustering approach to separate them effectively. Although the implicit grouping effect allows Instance-CL to attain better Accuracy and NMI scores, the resulting clusters are less apart from each other, and each cluster is more dispersed, as indicated by the smaller inter-cluster distance and larger intra-cluster distance. This result is unsurprising since Instance-CL only focuses on instance discrimination, which often leads to a more dispersed embedding space. In contrast, we leverage the strengths of both Clustering and Instance-CL to complement each other. Consequently, Figure 6 shows that ISCST yields better-separated clusters, with each cluster less dispersed. This effect is better than that of the previous SCCL clustering model because our ISCST model has a smaaller intra-cluster distance and a larger inter-cluster distance. Moreover, in terms of clustering metrics such as Accuracy and NMI, it also surpasses the previous SCCL clustering model as the number of iterations increases.

VI. CONCLUSION

We propose ISCST, a novel framework that leverages instance-wise contrastive learning to enhance unsupervised clustering with sentence-transformers. Evaluated across eight benchmark short-text datasets, our model substantially outperforms state-of-the-art methods. Ablation studies validate the effectiveness of integrating bottom-up instance discrimination with top-down clustering, generating high-quality clusters with improved intra- and inter-cluster distances. While evaluated on the short texts domain, our generic framework applies to various text clustering problems. Future work will explore data

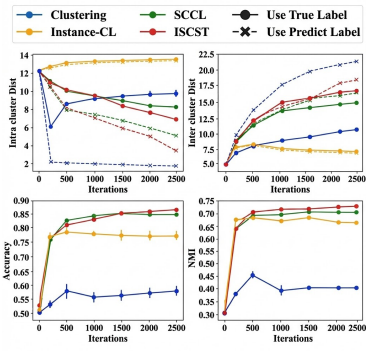

Fig. 6. Cluster-level evaluation on SearchSnippets. Each plot is summarized over ten random runs.

mixing strategies [33] to address natural language's discrete nature and augmentation challenges.

REFERENCES

[1] James MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, vol. 1, p. 281–297, 1967.

[2] Gilles Celeux and G´erard Govaert, "Gaussian parsimonious clustering models," Pattern Recognition, vol. 28, no. 5, pp. 781–793, 1995.

[3] Junyuan Xie, Ross Girshick, and Ali Farhadi, "Unsupervised deep embedding for clustering analysis," in International conference on machine learning, p. 478–487, 2016.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), p. 4171–4186, 2019.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," Computer Science, Linguistics, 2018.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint arXiv: 1910.13461, 2019.

[7] Richard Zhang, Phillip Isola, and Alexei A Efros, "Colorful image colorization," in European conference on computer vision, Springer, p. 649–666, 2016.

[8] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," arXiv preprint arXiv: 1803.07728, 2018.

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros, "Unsupervised visual representation learning by context prediction," in Proceedings of the IEEE international conference on computer vision, p. 1422–1430, 2015.

[10] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2019a.

[11] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin, "Unsupervised feature learning via nonparametric instance discrimination," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 3733–3742, 2018.

[12] Hwi-Gang Kim, Seongjoo Lee, and Sunghyon Kyeong, "Discovering hot topics using twitter streaming data social topic detection and geographic clustering," In 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), pages 1215–1220, IEEE, 2013.

[13] Christos Bouras and Vassilis Tsogkas, "Improving news articles recommendations via user clustering," International Journal of Machine Learning and Cybernetics, 8(1):223–237, 2017.

[14] Marc M Sebrechts, John V Cugini, Sharon J Laskowski, Joanna Vasilakis, and Michael S Miller, "Visualization of search results: a comparative evaluation of text, 2d, and 3d interfaces," In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 3–10, 1999.

[15] Suzanna Becker and Geoffrey E Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," Nature, vol. 355, no. 6356, pp. 161–163, 1992.

[16] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in Advances in Neural Information Processing Systems, pp. 766–774, Curran Associates, Inc., 2014.

[17] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019a.

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," arXiv preprint arXiv: 2004.11362, 2020.

[19] Amir Hadifar, Lucas Sterckx, Thomas Demeester, and Chris Develder, "A self-training approach for short text clustering," in Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), p. 194–199, 2019.

[20] Sanjeev Arora, Yingyu Liang, and Tengyu Ma, "A simple but tough-to-beat baseline for sentence embeddings," in International Conference on Learning Representations, 2017.

[21] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," arXiv preprint arXiv: 1802.05365, 2018.

[22] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[23] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," in NIPS-W, 2017.

[24] Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao, "Self-taught convolutional neural networks for short text clustering," Neural Networks, vol.88, no. 88, pp. 22–31, 2017.

[25] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," arXiv preprint arXiv: 2005.05909, 2020.

[26] Sosuke Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," arXiv preprint arXiv: 1805.06201, 2018.

[27] Chen Li, Xiaoguang Yu, Shuangyong Song, Jia Wang, Bo Zou, and Xiaodong He, "Simctc: A simple contrast learning method of text clustering (student abstract)," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 11, pp. 12997–12998, Jun. 2022.

[28] Md Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, and Evangelos Milios, "Enhancement of short text clustering by iterative classification," arXiv preprint arXiv: 2001.11631, 2020.

[29] Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang, "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, June 2021, pp. 5419–5430, Association for Computational Linguistics, 2021.

[30] Ruihui Li and Hongbin Wang, "Clustering of short texts based on dynamic adjustment for contrastive learning," IEEE Access, vol. 10, pp. 76069–76078, 2022.

[31] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi, "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, p. 91–100, 2008.

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

[33] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," arXiv preprint arXiv: 1710.09412, 2017b.