Supporting Clustering with Contrastive Learning

Dejiao Zhang[1]   **Feng Nan**[1]   **Xiaokai Wei**[1]   **Shang-Wen Li**[1]   **Henghui Zhu**[1]
Kathleen McKeown[1,2]   **Ramesh Nallapati**[1]   **Andrew O. Arnold**[1]   **Bing Xiang**[1]
[1]AWS AI   [2]Columbia University
`dejiaoz,nanfen,xiaokaiw,shangwel,henghui`
`mckeownk,rnallapa,anarnld,bxiang@amazon.com`

Abstract

Unsupervised clustering aims at discovering the semantic categories of data according to some distance measured in the representation space. However, different categories often overlap with each other in the representation space at the beginning of the learning process, which poses a significant challenge for distance-based clustering in achieving good separation between different categories. To this end, we propose Supporting Clustering with Contrastive Learning (SCCL) – a novel framework to leverage contrastive learning to promote better separation. We assess the performance of SCCL on short text clustering and show that SCCL significantly advances the state-of-the-art results on most benchmark datasets with $3\%-11\%$ improvement on Accuracy and $4\% - 15\%$ improvement on Normalized Mutual Information. Furthermore, our quantitative analysis demonstrates the effectiveness of SCCL in leveraging the strengths of both bottom-up instance discrimination and top-down clustering to achieve better intra-cluster and inter-cluster distances when evaluated with the ground truth cluster labels[1].

1 Introduction

Clustering, one of the most fundamental challenges in unsupervised learning, has been widely studied for decades. Long established clustering methods such as K-means (MacQueen et al., 1967; Lloyd, 1982) and Gaussian Mixture Models (Celeux and Govaert, 1995) rely on distance measured in the data space, which tends to be ineffective for high-dimensional data. On the other hand, deep neural networks are gaining momentum as an effective way to map data to a low dimensional and hopefully better separable representation space.

Many recent research efforts focus on integrating clustering with deep representation learning

---

[1]We plan to open source our implementation. Please visit `https://arxiv.org/abs/2103.12953` for the release updates.
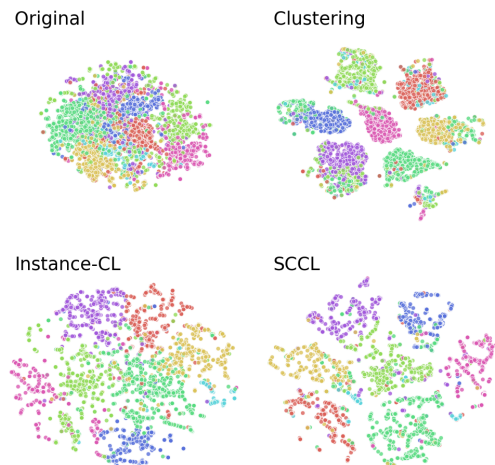

Figure 1: TSNE visualization of the embedding space learned on SearchSnippets using Sentence Transformer (Reimers and Gurevych, 2019a) as backbone. Each color indicates a ground truth semantic category.

by optimizing a clustering objective defined in the representation space (Xie et al., 2016; Jiang et al., 2016; Zhang et al., 2017a; Shaham et al., 2018). Despite promising improvements, the clustering performance is still inadequate, especially in the presence of complex data with a large number of clusters. 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.

On the other hand, Instance-wise Contrastive Learning (Instance-CL) (Wu et al., 2018; Bachman et al., 2019; He et al., 2020; Chen et al., 2020a,b) 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. Essen-

tially, Instance-CL disperses different instances apart while implicitly bringing similar instances together to some extent (see Figure 1). This beneficial property can be leveraged to support clustering by scattering apart the overlapped categories. Then clustering, thereby better separates different clusters while tightening each cluster by explicitly bringing samples in that cluster together.

To this end, we propose Supporting Clustering with Contrastive Learning (SCCL) by jointly optimizing a top-down clustering loss with a bottom-up instance-wise contrastive loss. We assess the performance of SCCL on short text clustering, which has become increasingly important due to the popularity of social media such as Twitter and Instagram. It benefits many real-world applications, including topic discovery (Kim et al., 2013), recommendation (Bouras and Tsogkas, 2017), and visualization (Sebrechts et al., 1999). However, the weak signal caused by noise and sparsity poses a significant challenge for clustering short texts. Although some improvement has been achieved by leveraging shallow neural networks to enrich the representations (Xu et al., 2017; Hadifar et al., 2019), there is still large room for improvement.

We address this challenge with our SCCL model. Our main contributions are the following:

- We propose a novel end-to-end framework for unsupervised clustering, which advances the state-of-the-art results on various short text clustering datasets by a large margin. Furthermore, our model is much simpler than the existing deep neural network based short text clustering approaches that often require multi-stage independent training.

- We provide in-depth analysis and demonstrate how SCCL effectively combines the top-down clustering with the bottom-up instance-wise contrastive learning to achieve better inter-cluster distance and intra-cluster distance.

- We explore various text augmentation techniques for SCCL, showing that, unlike the image domain (Chen et al., 2020a), using composition of augmentations is not always beneficial in the text domain.

2 Related Work

Self-supervised learning   Self-supervised learning has recently become prominent in providing effective representations for many downstream tasks.

Early work focuses on solving different artificially designed pretext tasks, such as predicting masked tokens (Devlin et al., 2019), generating future tokens (Radford et al., 2018), or denoising corrupted tokens (Lewis et al., 2019) for textual data, and predicting colorization (Zhang et al., 2016), rotation (Gidaris et al., 2018), or relative patch position (Doersch et al., 2015) for image data. Nevertheless, the resulting representations are tailored to the specific pretext tasks with limited generalization.

Many recent successes are largely driven by instance-wise contrastive learning. Inspired by the pioneering work of Becker and Hinton (1992); Bromley et al. (1994), Instance-CL treats each data instance and its augmentations as an independent class and tries to pull together the representations within each class while pushing apart different classes (Dosovitskiy et al., 2014; Oord et al., 2018; Bachman et al., 2019; He et al., 2020; Chen et al., 2020a,b). Consequently, different instances are well-separated in the learned embedding space with local invariance being preserved for each instance.

Although Instance-CL may implicitly group similar instances together (Wu et al., 2018), it pushes representations apart as long as they are from different original instances, regardless of their semantic similarities. Thereby, the implicit grouping effect of Instance-CL is less stable and more data-dependent, giving rise to worse representations in some cases (Khosla et al., 2020; Li et al., 2020; Purushwalkam and Gupta, 2020).

Short Text Clustering   Compared with the general text clustering problem, short text clustering comes with its own challenge due to the weak signal contained in each instance. In this scenario, BoW and TF-IDF often yield very sparse representation vectors that lack expressive ability. To remedy this issue, some early work leverages neural networks to enrich the representations (Xu et al., 2017; Hadifar et al., 2019), where word embeddings (Mikolov et al., 2013b; Arora et al., 2017) are adopted to further enhance the performance.

However, the above approaches divide the learning process into multiple stages, each requiring independent optimization. On the other hand, despite the tremendous successes achieved by contextualized word embeddings (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2018; Reimers and Gurevych, 2019b), they have been left largely unexplored for short text clustering. In this work, we leverage the pretrained transformer as the back-
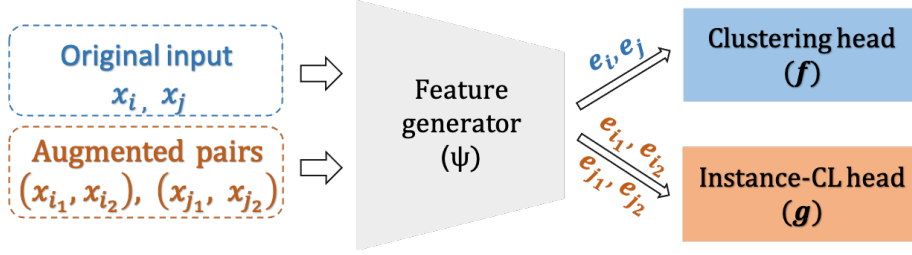
Figure 2: Training framework SCCL. During training, we jointly optimize a clustering loss over the original data instances and an instance-wise contrastive loss over the associated augmented pairs.

bone, which is optimized in an end-to-end fashion. As demonstrated in Section 4, we advance the state-of-the-art results on most benchmark datasets with $3\% - 11\%$ improvement on Accuracy and $4\% - 15\%$ improvement on NMI.

3 Model

We aim at developing a joint model that leverages the beneficial properties of Instance-CL to improve unsupervised clustering. As illustrated in Figure 2, our model consists of three components. A neural network $\psi(\cdot)$ first maps the input data to the representation space, which is then followed by two different heads $g(\cdot)$ and $f(\cdot)$ where the contrastive loss and the clustering loss are applied, respectively. Please refer to Section 4 for details.

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 a pair of augmentations for each data instance in $\mathcal{B}$, yielding an augmented batch $\mathcal{B}^a$ with size $2M$, denoted as $\mathcal{B}^a = \{\tilde{x}_i\}_{i=1}^{2M}$.

3.1 Instance-wise Contrastive Learning

For each minibatch $\mathcal{B}$, the Instance-CL loss is defined on the augmented pairs in $\mathcal{B}^a$. Let $i^1 \in \{1, \ldots, 2M\}$ denote the index of an arbitrary instance in augmented set $\mathcal{B}^a$, and let $i^2 \in \{1, \ldots, 2M\}$ be the index of the other instance in $\mathcal{B}^a$ augmented from the same instance in the original set $\mathcal{B}$. We refer to $\tilde{x}_{i^1}, \tilde{x}_{i^2} \in \mathcal{B}^a$ as a *positive* pair, while treating the other $2M$-2 examples in $\mathcal{B}^a$ as *negative* instances regarding this positive pair. Let $\tilde{z}_{i^1}$ and $\tilde{z}_{i^2}$ be the corresponding outputs of the head g, *i.e.,* $\tilde{z}_j = g(\psi(\tilde{x}_j)), j = i^1, i^2$. Then for $\tilde{x}_{i^1}$, we try to separate $\tilde{x}_{i^2}$ apart from all negative instances in $\mathcal{B}^a$ by minimizing the following

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

(1)

Here $\mathbb{1}_{j \neq i^1}$ is an indicator function and τ denotes the temperature parameter which we set as 0.5. Following Chen et al. (2020a), we choose $\text{sim}(\cdot)$ as the dot product between a pair of normalized outputs, *i.e.,* $\text{sim}(\tilde{z}_i, \tilde{z}_j) = \tilde{z}_i^T \tilde{z}_j / \|\tilde{z}_i\|_2 \|\tilde{z}_j\|_2$.

The Instance-CL loss is then averaged over all instances in $\mathcal{B}^a$,

$$\mathcal{L}_{\text{Instance-CL}} = \sum_{i=1}^{2M} \ell_i^I / 2M \; .$$

(2)

To explore the above contrastive loss in the text domain, we explore three different augmentation strategies in Section 4.3.1, where we find contextual augmenter (Kobayashi, 2018; Ma, 2019) consistently performs better than the other two.

3.2 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 tries to bring together instances from the same semantic category together. Suppose our data consists of K semantic categories, and each category is characterized by its centroid in the representation space, denoted as $\mu_k, k \in \{1, \ldots, K\}$. Let $e_j = \psi(x_j)$ denote the representation of instance x_j in the original set $\mathcal{B}$. Following Maaten and Hinton (2008), we use the Student's t-distribution to compute the probability of assigning x_j to the kth cluster,

$$q_{jk} = \frac{\left(1 + \|e_j - \mu_k\|_2^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K} \left(1 + \|e_j - \mu_{k'}\|_2^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \; .$$

(3)

Here α denotes the degree of freedom of the Student's t-distribution. Without explicit mention, we follow Maaten and Hinton (2008) by setting $\alpha = 1$ in this paper.

We use a linear layer, *i.e.,* the clustering head in Figure 2, to approximate the centroids of each

5421

cluster, and we iteratively refine it by leveraging an auxiliary distribution proposed by Xie et al. (2016). Specifically, let p_{jk} denote the auxiliary probability defined as

$$p_{jk} = \frac{q_{jk}^2/f_k}{\sum_{k'} q_{jk'}^2/f_{k'}} \ . \tag{4}$$

Here $f_k = \sum_{j=1}^M q_{jk}, k = 1, \ldots, K$ can be interpreted as the soft cluster frequencies approximated within a minibatch. This target distribution first sharpens the soft-assignment probability q_{jk} by raising it to the second power, and then normalizes it by the associated cluster frequency. By doing so, we encourage learning from high confidence cluster assignments and simultaneously combating the bias caused by imbalanced clusters.

We push the cluster assignment probability towards the target distribution by optimizing the KL divergence between them,

$$\ell_j^C = \mathrm{KL}\left[p_j \| q_j\right] = \sum_{k=1}^K p_{jk} \log \frac{p_{jk}}{q_{jk}} \ . \tag{5}$$

The clustering objective is then followed as

$$\mathcal{L}_{\mathrm{Cluster}} = \sum_{j=1}^M \ell_j^C/M \tag{6}$$

This clustering loss is first proposed in Xie et al. (2016) and later adopted by Hadifar et al. (2019) for short text clustering. However, they both require expensive layer-wise pretraining of the neural network, and update the target distribution (Eq (4)) through carefully chosen intervals that often vary across datasets. In contrast, we simplify the learning process to end-to-end training with the target distribution being updated per iteration.

Overall objective In summary, our overall objective is,

$$\mathcal{L} = \mathcal{L}_{\mathrm{Instance\text{-}CL}} + \eta \mathcal{L}_{\mathrm{Cluster}}$$
$$= \sum_{j=1}^M \ell_j^C/M + \eta \sum_{i=1}^{2M} \ell_i^I/2M \ . \tag{7}$$

ℓ_j^C and ℓ_i^I are defined in Eq (5) and Eq (2), respectively. η balances between the contrastive loss and the clustering loss of SCCL, which we set as 10 in Section 4 for simplicity. Also noted that, the clustering loss is optimized over the original data only. Alternatively, we can also leverage the augmented data to enforce local consistency of the cluster assignments for each instance. We discuss this further in Appendix A.3.

4 Numerical Results

Implementation We implement our model in PyTorch (Paszke et al., 2017) with the Sentence Transformer library (Reimers and Gurevych, 2019a). We choose *distilbert-base-nli-stsb-mean-tokens* as the backbone, followed by a linear clustering head (f) of size $768 \times K$ with K indicating the number of clusters. For the contrastive loss, we optimize an MLP (g) with one hidden layer of size 768, and output vectors of size 128. Figure 2 provides an illustration of our model. The detailed experimental setup is provided in Appendix A.1. We, as in the previous work Xu et al. (2017); Hadifar et al. (2019); Rakib et al. (2020), adopt Accuracy (ACC) and Normalized Mutual Information (NMI) to evaluate different approaches.

Datasets We assess the performance of the proposed SCCL model on eight benchmark datasets for short text clustering. Table 2 provides an overview of the main statistics, and the details of each dataset are as follows.

- **SearchSnippets** is extracted from web search snippets, which contains 12,340 snippets associated with 8 groups Phan et al. (2008).

- **StackOverflow** is a subset of the challenge data published by Kaggle[2], where 20,000 question titles associated with 20 different categories are selected by Xu et al. (2017).

- **Biomedical** is a subset of the PubMed data distributed by BioASQ[3], where 20,000 paper titles from 20 groups are randomly selected by Xu et al. (2017).

- **AgNews** is a subset of news titles (Zhang and LeCun, 2015), which contains 4 topics selected by Rakib et al. (2020).

- **Tweet** consists of 2,472 tweets with 89 categories (Yin and Wang, 2016).

- **GoogleNews** contains titles and snippets of 11,109 news articles related to 152 events (Yin and Wang, 2016). Following (Rakib et al., 2020), we name the full dataset as *GoogleNews-TS*, and *GoogleNews-T* and *GoogleNews-S* are obtained by extracting the titles and the snippets, respectively.

---

[2]https://www.kaggle.com/c/predict-closed-questions-on-stackoverflow/download/train.zip
[3]http://participants-area.bioasq.org

	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**
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

	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	-	-	-	-	-	-	-	-
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**

Table 1: Clustering results on eight short text datasets. Our results are averaged over five random runs.

For each dataset, we use *Contextual Augmenter* (Kobayashi, 2018; Ma, 2019) to obtain the augmentation set, as it consistently outperforms the other options explored in Section 4.3.1.

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

Table 2: Dataset statistics. $|V|$: the vocabulary size; N^D: number of short text documents; Len: average number of words in each document; N^C number of clusters; L/S: the ratio of the size of the largest cluster to that of the smallest cluster.

4.1 Comparison with State-of-the-art

We first demonstrate that our model can achieve state-of-the-art or highly competitive performance on short text clustering. For comparison, we consider the following baselines.

- **STCC** (Xu et al., 2017) consists of three independent stages. For each dataset, it first pretrains a word embedding on a large in-domain corpus using the Word2Vec method (Mikolov et al., 2013a). A convolutional neural network is then optimized to further enrich the representations that are fed into K-means for the final stage clustering.

- **Self-Train** (Hadifar et al., 2019) enhances the pretrained word embeddings in Xu et al. (2017) using SIF (Arora et al., 2017). Following Xie et al. (2016), it adopts an autoencoder obtained by layer-wise pretraining (Van Der Maaten, 2009), which is then further tuned with a clustering objective same as that in Section 3.2. Both Xie et al. (2016) and Hadifar et al. (2019) update the target distribution through carefully chosen intervals that vary across datasets, while we update it per iteration yet still achieve significant improvement.

- **HAC-SD** (Rakib et al., 2020)[4] applies hierarchical agglomerative clustering on top of a sparse pairwise similarity matrix obtained by zeroing-out similarity scores lower than a chosen threshold value.

- **BoW** & **TF-IDF** are evaluated by applying K-means on top of the associated features with dimension being 1500.

To demonstrate that our model is robust against the noisy input that often poses a significant chal-

---

[4]They further boost the performance via an iterative classification trained with high-confidence pseudo labels extracted after each round of clustering. Since the iterative classification strategy is orthogonal to the clustering algorithms, we only evaluate against with their proposed clustering algorithm for fair comparison.
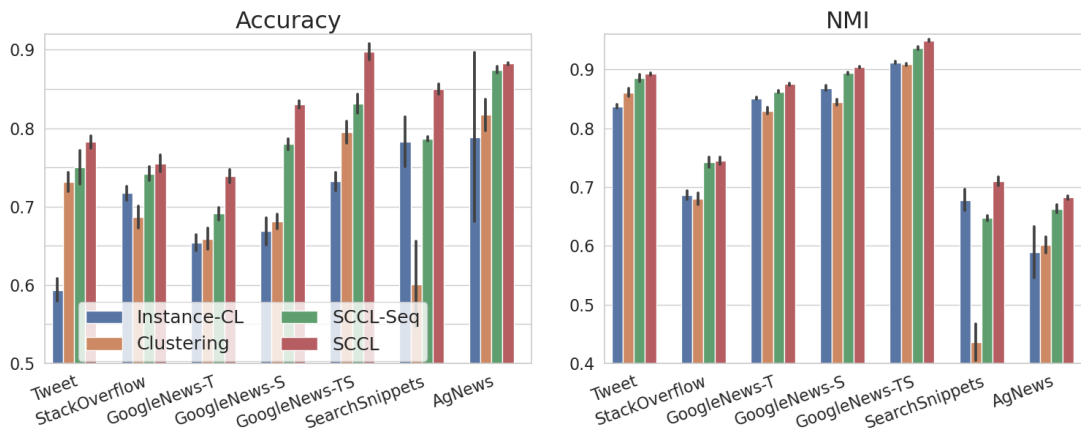
Figure 3: Ablation study of SCCL. In SCCL-Seq, we first train the model using Instance-CL, and then optimize the clustering objective. We exclude Biomedical for better visualization, full plot can be found in Appendix A.4.

lenge for short text clustering, we do not apply any pre-processing procedures on any of the eight datasets. In contrast, all baselines except BoW and TF-IDF considered in this paper either pre-processed the Biomedical dataset (Xu et al., 2017; Hadifar et al., 2019) or all eight datasets by removing the stop words, punctuation, and converting the text to lower case (Rakib et al., 2020).

We report the comparison results in Table 1. Our SCCL model outperforms all baselines by a large margin on most datasets. Although we are lagging behind Hadifar et al. (2019) on Biomedical, SCCL still shows great promise considering the fact that Biomedical is much less related to the general domains on which the transformers are pretrained. In contrast, Hadifar et al. (2019) learn the word embeddings on a large in-domain biomedical corpus, followed by a layer-wise pretrained autoencoder to further enrich the representations.

Rakib et al. (2020) also shows better Accuracy on Tweet and GoogleNews-T, for which we hypothesize two reasons. First, both GoogleNews and Tweet have fewer training examples with much more clusters. Thereby, it's challenging for instance-wise contrast learning to manifest its advantages, which often requires a large training dataset. Second, as implied by the clustering performance evaluated on BoW and TF-IDF, clustering GoogleNews and Tweet is less challenging than clustering the other four datasets. Hence, by applying agglomerative clustering on the carefully selected pairwise similarities of the preprocessed data, Rakib et al. (2020) can achieve good performance, especially when the text instances are very short, *i.e.,* Tweet and GoogleNews-T. We also high-

light the scalability of our model to large scale data, whereas agglomerative clustering often suffers from high computation complexity. We discuss this further in Appendix A.5.

4.2 Ablation Study

To better validate our model, we run ablations in this section. For illustration, we name the clustering component described in Section 3.2 as Clustering. Besides Instance-CL and Clustering, we also evaluate SCCL against its sequential version (SCCL-Seq) where we first train the model with Instance-CL, and then optimize it with Clustering.

As shown in Figure 3, Instance-CL also groups semantically similar instances together. However, this grouping effect is implicit and data-dependent. In contrast, SCCL consistently outperforms both Instance-CL and Clustering by a large margin. Furthermore, SCCL also achieves better performance than its sequential version, SCCL-Seq. The result validates the effectiveness and importance of the proposed joint optimization framework in leveraging the strengths of both Instance-CL and Clustering to compliment each other.

4.2.1 SCCL leads to better separated and less dispersed clusters

To further investigate what enables the better performance of SCCL, we track both the intra-cluster distance and the inter-cluster distance evaluated 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 4, we report each type

Dataset	Accuracy			NMI		
	WNet	**Para**	**Ctxt**	**WNet**	**Para**	**Ctxt**
AgNews	86.6	86.5	**88.2**	66.0	65.2	**68.2**
SearchSnippets	78.1	83.7	**85.0**	61.9	68.1	**71.0**
StackOverflow	69.1	73.3	**75.5**	69.9	72.7	**74.5**
Biomedical	42.8	43.0	**46.2**	38.0	39.5	**41.5**
GoooglenewsTS	82.1	83.5	**89.8**	92.1	92.9	**94.9**
GooglenewsS	73.0	75.3	**83.1**	86.4	87.4	**90.4**
GooglenewsT	66.3	67.5	**73.9**	83.4	83.6	**87.5**
Tweet	70.6	73.7	**78.2**	86.2	86.4	**89.2**

Table 3: Results of SCCL evaluated with different augmentation techniques: WordNet augmenter (**WNet**), paraphrase via back translation (**Para**), and contextual augmenter (**Ctxt**). Each technique is detailed in Section 4.3.1.

of distance with its mean value obtained by averaging over all clusters, where the clusters are defined either regarding the ground truth labels (solid lines) or the labels predicted by the model (dashed lines).
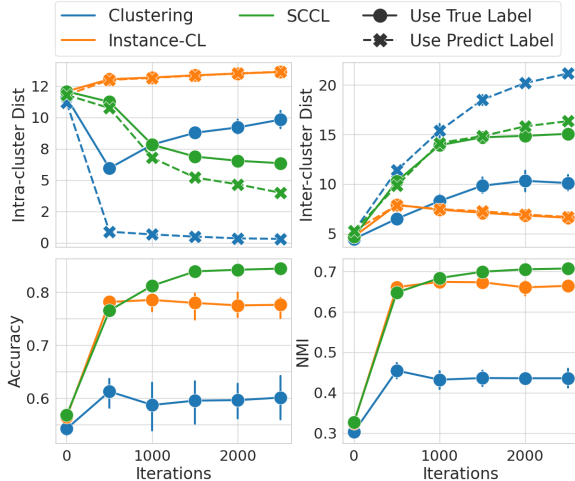

Figure 4: Cluster-level evaluation on SearchSnippets. Each plot is summarized over five random runs.

Figure 4 shows Clustering achieves smaller intra-cluster distance and larger inter-cluster distance when evaluated on the predicted clusters. It demonstrates the ability of Clustering to tight each self-learned cluster and separate different clusters apart. However, we observe the opposite when evaluated on the ground truth clusters, along with poor Accuracy and NMI scores. One possible explanation is, data from different ground-truth clusters often have significant overlap in the embedding space before clustering starts (see upper left plot in Figure 1), which makes it hard for our distance-based clustering approach to separate them apart effectively.

Although the implicit grouping effect allows Instance-CL attains 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 compliment each other. Consequently, Figure 4 shows SCCL leads to better separated clusters with each cluster being less dispersed.

4.3 Data Augmentation

4.3.1 Exploration of Data Augmentations

To study the impact of data augmentation, we explore three different unsupervised text augmentations: (1) *WordNet Augmenter*[5] transforms an input text by replacing its words with WordNet synonyms (Morris et al., 2020; Ren et al., 2019). (2) *Contextual Augmenter*[6] leverages the pretrained transformers to find top-n suitable words of the input text for insertion or substitution (Kobayashi, 2018; Ma, 2019). We augment the data via word substitution, and we choose Bertbase and Roberta to generate the augmented pairs. (3) *Paraphrase via back translation*[7] generates paraphrases of the input text by first translating it to another language (French) and then back to English. When translating back to English, we used the mixture of experts model (Shen et al., 2019) to generate ten candidate paraphrases per input to increase diversity.

For both *WordNet Augmenter* and *Contextual Augmenter*, we try three different settings by choosing the word substitution ratio of each text instance

---

[5] https://github.com/QData/TextAttack
[6] https://github.com/makcedward/nlpaug
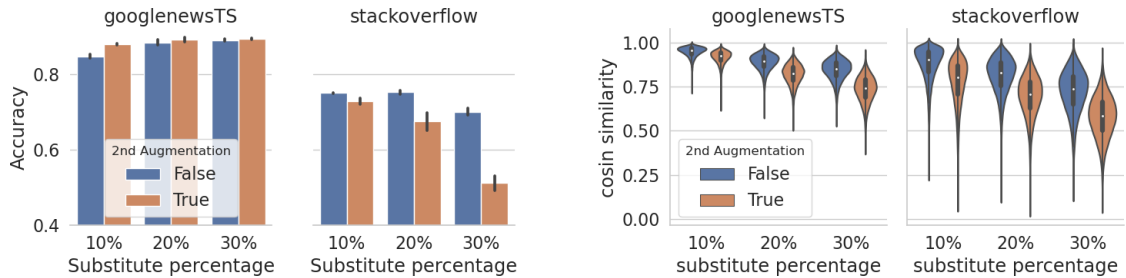[7] https://github.com/pytorch/fairseq/tree/master/examples/paraphraser

Figure 5: Impact of using composition of data augmentations. Either only Contextual Augmenter (blue) is used, or Contextual Augmenter and CharSwap Augmenter are applied sequentially (orange). **(Left)** 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. **(Right)** Distribution of the cosine similarity between the representations of each original text and its augmented pair at the beginning of training.

to 10%, 20%, and 30%, respectively. As for *Paraphrase via back translation*, we compute the BLEU score between each text instance and its ten candidate paraphrases. We then select three pairs, achieving the highest, medium, and lowest BLEU scores, from the ten condidates of each instance. The best results[8] of each augmentation technique are summarized in Table 3, where *Contexual Augmenter* substantially outperforms the other two. We conjecture that this is due to both *Contextual Augmenter* and SCCL leverage the pretrained transformers as backbones, which allows *Contextual Augmenter* to generate more informative augmentations.

4.3.2 Composition of Data Augmentations

Figure 5 shows the impact of using composition of data augmentations, in which we explored *Contextual Augmenter* and *CharSwap Augmenter*[9] (Morris et al., 2020). As we can see, using composition of data augmentations does boost the performance of SCCL on GoogleNews-TS where the average number of words in each text instance is 28 (see Table 2). 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 composition of data augmentations is crucial for contrastive learning to attain good performance. Possible explanations is that generating high-quality augmentations for textual data is more challenging, since changing a single word can invert the semantic meaning of the whole instance. This challenge is compounded when a second round of augmentation is applied on very

short text instances, *e.g.,* StackOverflow. We further demonstrate this in Figure 5 (right), where the augmented pairs of StackOverflow largely diverge from the original texts in the representation space after the second round of augmentation.

5 Conclusion

We have proposed a novel framework leveraging instance-wise contrastive learning to support unsupervised clustering. We thoroughly evaluate our model on eight benchmark short text clustering datasets, and show that our model either substantially outperforms or performs highly comparably to the state-of-the-art methods. Moreover, we conduct ablation studies to better validate the effectiveness of our model. We demonstrate that, by integrating the strengths of both bottom-up instance discrimination and top-down clustering, our model is capable of generating high-quality clusters with better intra-cluster and inter-clusters distances. Although we only evaluate our model on short text data, the proposed framework is generic and is expected to be effective for various kinds of text clustering problems.

In this work, we explored different data augmentation strategies with extensive comparisons. However, due to the discrete nature of natural language, designing effective transformations for textual data is more challenging compared to the counterparts in the computer vision domain. One promising direction is leveraging the data mixing strategies (Zhang et al., 2017b) to either obtain stronger augmentations (Kalantidis et al., 2020) or alleviate the heavy burden on data augmentation (Lee et al., 2020). We leave this as future work.

---

[8] Please refer to Appendix A.2 for details.
[9] A simple technique that augments text by substituting, deleting, inserting, and swapping adjacent characters

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545.

Suzanna Becker and Geoffrey E Hinton. 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163.

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

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744.

Gilles Celeux and Gérard Govaert. 1995. Gaussian parsimonious clustering models. *Pattern recognition*, 28(5):781–793.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

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

Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774.

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

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. 2016. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *arXiv preprint arXiv:2010.01028*.

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

Hwi-Gang Kim, Seongjoo Lee, and Sunghyon Kyeong. 2013. 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. 2020. i-mix: A strategy for regularizing contrastive representation learning. *arXiv preprint arXiv:2010.08887*.

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

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

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

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *arXiv preprint arXiv:2102.08473*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

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

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.

Senthil Purushwalkam and Abhinav Gupta. 2020. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. 2020. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. *arXiv preprint arXiv:2010.08670*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

Nils Reimers and Iryna Gurevych. 2019a. 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.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Marc M Sebrechts, John V Cugini, Sharon J Laskowski, Joanna Vasilakis, and Michael S Miller. 1999. 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.

Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. 2018. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. *International Conference on Machine Learning*.

Laurens Van Der Maaten. 2009. Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics*, pages 384–391.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487.

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.

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, Yifan Sun, Brian Eriksson, and Laura Balzano. 2017a. Deep unsupervised clustering using mixture of autoencoders. *arXiv preprint arXiv:1712.07788*.

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

Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.

A   Appendices

A.1   Experiment Setup

We use the Adam optimizer (Kingma and Ba, 2015) with batch size of 400. We use *distilbert-base-nli-stsb-mean-tokens* in the Sentence Transformers library (Reimers and Gurevych, 2019a) as the backbone, and we set the maximum input length to 32. We use a constant learning rate 5e-6 to optimize the backbone, while setting learning rate to 5e-4 to optimizing both the Clustering head and Instance-CL head. Same as Xie et al. (2016); Hadifar et al. (2019), we set $\alpha = 1$ for all datasets except Biomedical where we use $\alpha = 10$. As mentioned in Section 3, we set $\tau = 0.5$ for optimizing the constrastive loss. We tried different τ values in the range of $(0, 1]$ and found using $\tau = 0.5$ yields comparatively better yet stable performance across datasets for both Instance-CL and SCCL. For fair comparison between SCCL and its components or variants, we report the clustering performance for each of them by applying KMeans on the representations post the associated training processes.

A.2   Data Augmentation

Exploration of Data Augmentations   As mentioned in Section 4.3.1, we tried three different augmentation strengths for both *WordNet Augmenter*

and *Contextual Augmenter* by choosing the word substitution ratio of each text instance as 10%, 20%, and 30%, respectively. For each augmentation strength, we generate a pair of augmentations for each text instance. As for *Paraphrase via back translation*, we computed the BLEU score between each original instance and its ten candidate paraphrases. We then select three pairs, achieving the highest, medium, and lowest BLEU scores, from the ten condidates as the augmented data. For each augmentation method, we run SCCL on all three augmentation strengths independently and report the best result.

For both *WordNet Augmenter* and *Contextual Augmenter*, we observe that comparatively longer text instances, *i.e.,* those in AgNews, SearchSnippets, GoogleNewsTS, and GoogleNewsS, benefit from stronger augmentation. In contrast, *Paraphrase via back translation* shows better results when evaluated on the augmented pairs achieving the lowest BLEU scores with the original instance, *i.e.,* the pair achieving the two lowest scores among all ten condidate paraphrases for each text instance.

Building Effective Data Augmentations for NLP   As discussed in Section 4.3.2, using composition of data augmentations is not always beneficial for short text clustering. Because changing a single word can invert the meaning of the whole sentence, and the challenge is compounded when applied a second round data augmentation to short text data. However, we would hopefully cross the hurdle soon, as more effective approaches are keeping developed by the NLP community Qu et al. (2020); Giorgi et al. (2020); Meng et al. (2021).

A.3   Alternative Clustering Loss for SCCL

In the current form of SCCL, the clustering loss is optimized on the original dataset only. However, several alternatives could be considered, we discuss two options here to encourage further explorations.

Alternative 1.   Let j^1 and j^2 denote the indices of the augmented pair for the jth text instance in the original set, respectively. For the augmented instance j^1, we then push the cluster assignment probability towards the target distribution obtained by the other instance j^2, and vice versa. That is, we replace Eq (5) with the following

$$\ell_j^C = \text{KL}\left[p_{j^1}||q_{j^2}\right] + \text{KL}\left[p_{j^2}||q_{j^1}\right] \qquad (8)$$

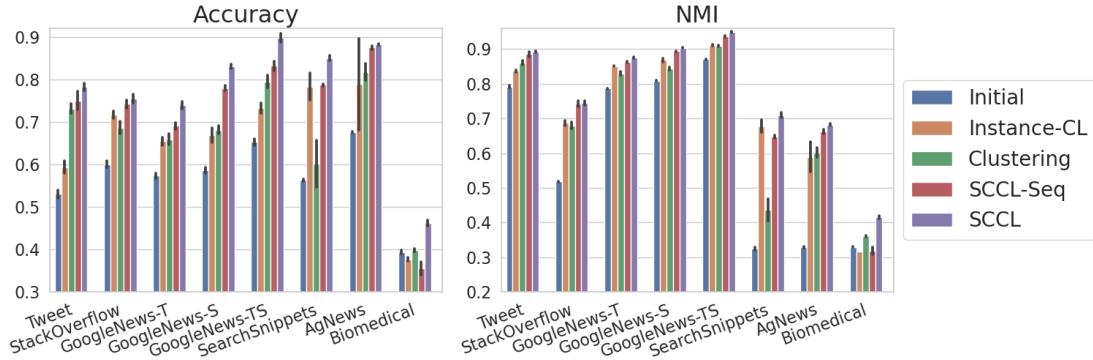Here p and q denote the target distribution and the cluster assignment probability defined in Eqs (3)

Figure 6: Ablation study of our proposed SCCL model. In SCCL-Seq, we first train the model using Instance-CL, and then optimize the clustering objective.

and (4), respectively.

Alternative 2. Let j^0 and j^1, j^2 denote the indices of the original text instance and its augmented pair, respectively. We then use the original instance as anchor, and push the cluster assignments of the augmented pair towards it by optimizing the following

$$\ell_j^C = \text{KL}\left[p_{j^0}||q_{j^1}\right] + \text{KL}\left[p_{j^0}||q_{j^2}\right] \quad (9)$$

Exploring (8) and (9) is out of the scope of this paper, however, it's worth trying when applying SCCL to solve different application problems. Especially considering that the above alternatives might lead to further performance improvement by jointly optimizing the instance-level and the cluster assignment level contrastive learning losses.

A.4 Supplement materials for ablation study

Figure 6 provides the full version of Figure 3 in Section 4.

A.5 Comparison with Rakib et al. (2020)

While Rakib et al. (2020) achieve better Accuracy on Tweet and GoogleNews-T, we highlight the scalability of our model to large scale data, whereas Rakib et al. (2020) depend on agglomerative clustering which often suffers from high computation complexity. Specifically, let N denote the number of training examples, and K denote the number of clusters. The HAC-SD method proposed by Rakib et al. (2020) first computes the pairwise similarity among all possible pairs of the data, and then sorts the N^2 similarity values so as to select the top N^2/K pairwise similarity as the input to the agglomerative clustering algorithm. Thereby, before clustering, HAC-SD could result in $O(N^2 \log N)$

time complexity, and $O(N^2/K)$ storage complexity. Moreover, the agglomerative clustering algorithm could require $O(N^2 \log(N/K))$ time complexity. Therefore, HAC-SD is less feasible in presence of large scale data. In contrast, SCCL performs standard stochastic optimization, the time complexity linearly scales with N since SCCL often requires $20 - 100$ epochs to converge, which is often much smaller than the number of data examples.