MIST: Mutual Information Maximization for Short Text Clustering

Anonymous ACL submission

Abstract

Short text clustering poses substantial challenges due to the limited amount of information provided by each text sample. Previous efforts based on dense representations are still inadequate as texts are not sufficiently segregated in the embedding space before clustering. Even though the state-of-the-art method utilizes contrastive learning to boost performance, the process of summarizing all local tokens to form a sequence representation for the whole text includes noise that may obscure limited key information. We propose **M**utual **I**nformation Maximization Framework for **S**hort **T**ext Clustering (MIST), which overcomes the information drown-out by including a mechanism to maximize the mutual information between representations on both sequence and token levels. Experimental results across eight standard short text datasets show that MIST outperforms the state-of-the-art method in terms of Accuracy or Normalized Mutual Information in most cases.

1 Introduction

Text clustering is a vital task for a wide range of downstream applications. It aims to partition texts into groups of similar categories in an unsupervised manner. The growth of social media, discussion forums, and news aggregator websites has led to a large number of short-length texts being produced daily. Therefore, clustering these short texts has become crucial for many real-world applications ranging from recommendation to text retrieval (Yohannes and Assabie, 2021).

In short texts, the most informative words and phrases of the text content usually appear only once. This exacerbates the sparsity problem, posing an additional hurdle for clustering short texts. Traditional methods, such as BoW and TF-IDF, provide relatively sparse representation vectors with limited descriptive power. Hence, they perform poorly when clustered using a standard distance-based clustering algorithm (Hadifar et al., 2019).

To address this problem, most recent methods (Xu et al., 2017; Hadifar et al., 2019; Yin et al., 2021) utilize deep neural networks to map high-dimensional data into meaningful dense representations in a lower-dimensional space and adopt a multi-stage scheme in which the clustering process is performed after learning feature representations. However, the clustering performance of these methods remains unsatisfactory as texts still have a lot of overlap among categories in the latent space before clustering (Zhang et al., 2021).

Alternatively, an end-to-end clustering scheme (Zhang et al., 2021; Xie et al., 2016) simultaneously optimizes representation learning and clustering objectives. To achieve desirable outcomes, Zhang et al. (2021) propose a method that employs contrastive representation learning, which has been successful in self-supervised learning and can help spread out overlapping categories, in order to obtain effective short text representations.

As shown in Zhang et al. (2021), improving representation is crucial for enhancing the clustering performance. Nevertheless, the contrastive learning method used in Zhang et al. (2021) only considers sequence-level embeddings that are formed by averaging all local tokens in each text instance, including uninformative noise. This could generate a representation in which spare yet informative terms used to describe the text content may be obscured by noise, potentially affecting the clustering performance. We consider the preservation of limited information in such a low signal-to-noise environment as a vital feature for short-text clustering. Addressing this gap will result in sequence representations that are more semantically representative and robust to noisy tokens in short texts.

In this paper, we introduce the **M**utual **I**nformation Maximization Framework for **S**hort **T**ext Clustering (MIST), a new multi-stage approach. We aim to improve representation learning stage for short text clustering using two contrastive learning objectives operating at the sequence and token levels. In particular, we apply the concept of mutual information (MI) maximization to facilitate us in comparing the semantic similarity between representations across the two hierarchical levels.

The crux of our method lies in integrating the *sequence-level* and *token-level* MI maximization objectives concurrently for the following purposes.

1. *Learning Distinct Text Representation:* The first learning objective maximizes MI between each positive pair at the sequence level;
2. *Informative Token Preservation:* The second objective is designed to enforce each text representation at the sequence level to extract local information shared across all its individual tokens by directly maximizing MI between them. This way, we mitigate the obscurity-by-noise problem and preserve limited key information in a weak signal environment.

The growth in the size of short text sequences may exacerbate a poor signal-to-noise ratio. To deal with short text samples with various signal-to-noise ratios, we additionally propose an *adaptive weighting function* that dynamically determines an appropriate ratio between the two objectives based on the length of the texts. To our knowledge, the method of combining two MI maximization objectives logically is presented for the first time. Note that the representations at different levels have a direct implication on one another, and the sequence representations are subsequently used in the clustering stage by applying the k-means algorithm.

We conduct extensive experimental studies over the eight standard benchmarks. MIST improves the clustering performance in terms of Accuracy and Normalized Mutual Information in most cases compared to the current state-of-the-art while using an identical configuration across all datasets. This demonstrates the generalizability of our method.

Our main contributions are outlined as follows: (1) We propose a novel representation learning technique for short text clustering through the integration of sequence-level and token-level MI maximization objectives. (2) To balance the two objectives, we introduce an adaptive weighting function. (3) Our ablation study provides a further demonstration of how different prioritization of the two MI objectives impacts the clustering performance across datasets of various text lengths; as text length increases, the preservation of limited local information becomes more significant.

2 Related Work

Short Text Clustering. There are several strategies to overcome the sparsity of short text representations. Some recent methods utilize a multi-stage architecture that breaks down the clustering framework into multiple stages; the clustering process is performed after learning feature representations. Xu et al. (2015, 2017) use a convolutional neural network to learn non-biased representations by fitting the output units with pretrained-binary codes from a dimensionality reduction method. Hadifar et al. (2019) utilize Smooth Inverse Frequency (Arora et al., 2017) to obtain weighted word embeddings. During training, they enrich discriminative features by tuning an autoencoder with soft clustering assignments. For the aforementioned works, the k-means clustering is employed on the trained representations to get the final clusters.

Another approach is to enhance the quality of the initial clustering with an iterative classification algorithm. Rakib et al. (2020) proposed the ECIC algorithm to detect and remove outliers in each iteration. Moreover, they make use of word embeddings by averaging them to represent each text, and combine the ECIC algorithm with hierarchical clustering. To boost the clustering quality further, Pugachev and Burtsev (2021) exploit deep sentence representations (Cer et al., 2018) and make modifications to the ECIC algorithm.

The recent state-of-the-art, SCCL (Zhang et al., 2021), leverages contrastive learning to encourage greater separation between overlapped categories in the original data space. By jointly optimizing a contrastive loss and a clustering objective (Reimers and Gurevych, 2019a), SCCL outperforms prior works and yields cutting-edge results. In addition, other constrastive learning methods have also been experimented on short text clustering, such as using entities for contrastive learning to provide supervision signals for their related sentences (Nishikawa et al., 2022), and using virtual augmentation for contrastive learning to circumvent the discrete nature of language (Zhang et al., 2022).

Moreover, a new technique for short text clustering is presented in Zheng et al. (2023); it comprises a pseudo-label generation module and a robust representation learning module. The former generates pseudo-labels, which are robust against the imbalance in data, as the supervision for the latter.

Self-Supervised Learning. Self-supervision has gained popularity and become a common technique

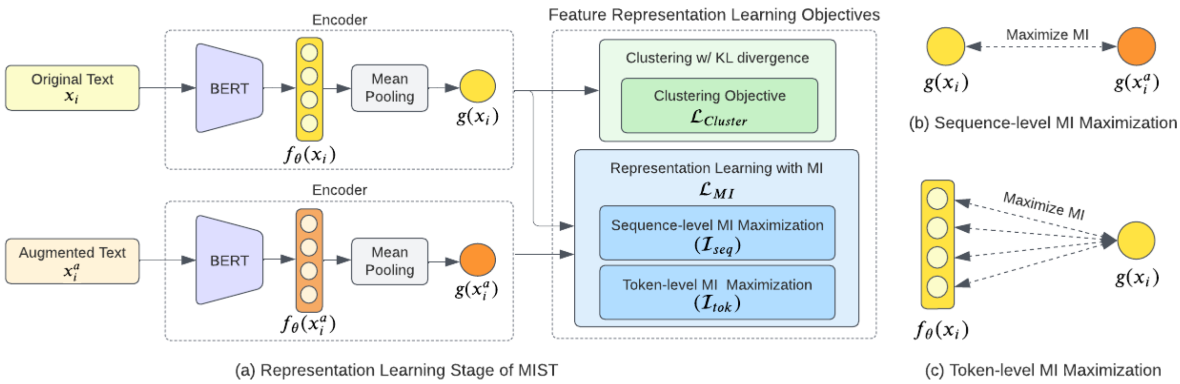

Figure 1: (a) Representation Learning Stage Overview. MIST considers all pairs of original text x_i, and its augmented version x_i^a as positive samples. MIST jointly optimizes the clustering objective $\mathcal{L}_{\text{Cluster}}$, and the MI objective $\mathcal{L}_{\text{MI}}$, which includes (b) a sequence-level MI maximization objective $\mathcal{I}_{\text{seq}}$ that maximize MI between representations at the sequence level (x_i and x_i^a), and (c) a token-level MI maximization objective $\mathcal{I}_{\text{tok}}$ that directly maximizes MI between a sequence representation(of both x_i and x_i^a) and its tokens ($f_\theta(x_i)$ and $f_\theta(x_i^a)$).

in unsupervised representation learning for a variety of downstream purposes(Chen et al., 2020; He et al., 2020; Caron et al., 2020; Grill et al., 2020).

Learning meaningful representations by estimating and maximizing MI is one of the prominent contrastive learning strategies. Its effectiveness has been demonstrated in both vision (Hjelm et al., 2019; Bachman et al., 2019; Sordoni et al., 2021) and text domains (Kong et al., 2020; Caron et al., 2020; Giorgi et al., 2021). Deep Infomax (DIM) (Hjelm et al., 2019) introduces global and local MI maximization objectives for learning image representations. However, each of these is implemented separately according to the task. The authors find success in optimizing local MI maximization objective by maximizing MI between local features and global features. Inspired by local Deep InfoMax, Zhang et al. (2020) proposes a sentence representation learning method that maximizes the MI between the sentence-level representation and its CNN-based n-gram contextual dependencies.

On the contrary, we integrate two MI maximization strategies concurrently to learn textual representations for various short text characteristics. We also introduce a generalized adaptive weighting function for effectively integrating both objectives.

3 Proposed Method: MIST

We propose a short text clustering framework consisting of two stages. First, we train a model using feature representation learning objectives as illustrated in Figure 1. Second, we apply the k-means algorithm on the trained representations at inference time to obtain the final clusters. This investigation focuses on improving the first stage.

The main idea of our solution lies in the learning objective function $\mathcal{L}$ that takes into account an MI objective $\mathcal{L}_{\text{MI}}$ and an unsupervised clustering objective $\mathcal{L}_{\text{Cluster}}$, which is used to enforce the encoder to capture categorical structure and provide a suitable representation space for clustering task.

$$\mathcal{L} = \beta\mathcal{L}_{\text{MI}} + \eta\mathcal{L}_{\text{Cluster}}, \tag{1}$$

where β and η represent the trade-off between $\mathcal{L}_{\text{MI}}$, and $\mathcal{L}_{\text{Cluster}}$. In our experiments, we set β to 1 and η to 2 to provide more weight to $\mathcal{L}_{\text{Cluster}}$.

In Section 3.1, we describe our main contribution, the MI maximization learning procedure, including (1) sequence-level and token-level MI maximization objectives; (2) an adaptive weighting function that is also incorporated to balance them. Section 3.2 presents the auxiliary clustering objective utilized in the learning stage.

3.1 Representation Learning with MI Maximization

Short texts are challenging to cluster due to the weak signal caused by noise. In the context of this study, short texts are recognized as those that are short in length and typically contain informal fragmental non-sentence structures, e.g., tweets and news snippets. One strategy to improve the clustering performance is to adopt *contrastive learning* to construct an embedding space that minimizes local invariance for each positive pair. However, a standard contrastive learning procedure, which is performed by contrasting between sequence representations (global features), may allow noise to *drown out* sparse but informative local-token embeddings (local features) when these tokens are mean-pooled to form a sequence representation. Consequently, optimizing solely contrastive learning at the sequence level is insufficient for learning representations in a weak signal environment.

3.1.1 Hierarchical MI Objective

In contrast to previous works on MI maximization learning, which utilized each MI objective separately, we incorporate the learning of both sequence and token representations into a single objective. This strategy offers two advantages: (1) it mitigates the problem of information drown-out by allowing individual tokens to participate in the MI maximization process; (2) it supports weight adjustment between these two MI levels to handle short text inputs with various signal-to-noise ratios.

Sequence-Token MI Maximization. According to Tian et al. (2020), contrastive learning is equivalent to maximizing the lower bound of MI between a sequence representation and its augmented version (positive). Intuitively, it reflects how much more precisely we can determine the representation given a positive compared to when we are unaware of the positive (Bachman et al., 2019). This principle enables us to incorporate an additional mechanism beyond the sequence-level objective.

We build our framework based on the MI maximization concept through the integration of two MI objectives. In this way, our model can effectively learn distinct short text representations using the *sequence-level* MI objective while simultaneously preserving local information using an additional objective. Specifically, the *token-level* MI objective helps alleviate the information obscurity from noise by maximizing the MI between each local token and its sequence representation. As a result, the overall learning objective $\mathcal{L}_{\text{MI}}$ consists of two components: (1) sequence-level MI maximization $\mathcal{I}_{\text{seq}}$, and (2) token-level MI maximization $\mathcal{I}_{\text{tok}}$, operating concurrently in a sequence-token hierarchy as shown in Figure 1.

$$\mathcal{L}_{\text{MI}} = -(1-\lambda)\mathcal{I}_{\text{seq}} - \lambda\mathcal{I}_{\text{tok}}, \tag{2}$$

where λ corresponds to the balancing weight for $\mathcal{I}_{\text{seq}}$ and $\mathcal{I}_{\text{tok}}$ objectives, which is defined in Eq.3.

Adaptive Weighting Function. According to our analysis, short text inputs vary in length across different datasets, ranging from *fragmental sequences* of 6 words to 28 words. Regarding signal-to-noise, larger sequences tend to contain a greater proportion of noise that does not provide useful semantics for the clustering step, whereas informative terms *usually still appear once*. This exacerbates the information drown-out problem due to a poor signal-to-noise ratio.

While other short text clustering techniques treat all text samples in the same fashion, we argue that different-length short texts should be handled differently. We propose an MI maximization strategy adaptable to text length so that our method can efficiently deal with short text instances containing varying signal-to-noise ratios, without the need for a hyperparameter search for any particular dataset. Since larger sequences necessitate more effort to preserve limited crucial information, we place more weight on the $\mathcal{I}_{\text{tok}}$ objective by encouraging λ to be larger as the total number of tokens in the text grows. Thus, our *generalized adaptive weighting function* (Eq.3) is introduced to assign the weight of λ depending on the average number of tokens in text samples for each minibatch of size N:

$$\lambda = \max\left(0, \left\lfloor \frac{0.1}{N}\sum_{i=1}^{N} l_i \right\rceil - 1\right) \times 0.1, \tag{3}$$

where l_i denotes the number of tokens in a text x_i and it is directly proportional to the text length.

In the representation learning stage, we randomly sample a minibatch $X^o = x_1^o, ..., x_N^o$ of N original texts with empirical probability distribution $\mathbb{P}$. Then, we generate an augmented version for each text to obtain an augmented batch $X^a = x_1^a, ..., x_N^a$, where X^o and X^a are of identical size. The encoder f_θ, a pretrained transformer network, encodes an input text x into a sequence of contextualized token embeddings with length l, $f_\theta(x) := \{f_\theta^{(i)}(x) \in \mathbb{R}^d\}_{i=1}^{l}$, where i is the token index and d is the number of dimension. These token representations are then mean pooled $m(f_\theta(x))$ to generate a sequence representation denoted as $g(x) = m(f_\theta(x)) \in \mathbb{R}^d$.

3.1.2 Computing the Sequence-level MI.

This learning objective, $\mathcal{I}_{seq}$, aims to learn distinct text representations through the maximization of MI between the original sample and its augmented version at the sequence level. By treating each original text $g(x^o)$ and its augmentation $g(x^a)$ as positive pairs, the $\mathcal{I}_{seq}$ objective is defined as:

$$\mathcal{I}_{seq} = \frac{1}{N}\left(\sum_{x\in X} \widehat{\mathcal{I}}^{JSD}(g(x^o); g(x^a))\right) \tag{4}$$

We adopt a Jensen-Shannon estimator (Nowozin et al., 2016; Hjelm et al., 2019) to estimate a lower bound of MI, $\widehat{\mathcal{I}}_\theta^{JSD}$:

$$\widehat{\mathcal{I}}_{\theta}^{JSD}(g(x^o); g(x^a)) := E_{\mathbb{P}}\left[-sp(-g(x^o) \cdot g(x^a)).\right] - E_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[sp(g(x^o) \cdot g(\tilde{x}^a))\right], \quad (5)$$

where $\tilde{x}^a$ is a negative augmented textual input sampled from distribution $\tilde{\mathbb{P}} = \mathbb{P}$, and $sp(z) = \log(1 + e^z)$ is the softplus function.

3.1.3 Computing the Token-level MI

In contrast to Zhang et al. (2020), we constrain the sequence representation containing high MI with each token to preserve limited local information in short texts— by maximizing MI between the sequence representation and its token representations directly— instead of its local contextual n-gram embeddings. In particular, we attempt to maximize the average MI between a sequence representation and all its token representations while minimizing MI with the tokens of other texts. We define $\mathcal{I}_{tok}$ for each minibatch as

$$\mathcal{I}_{tok} = \frac{1}{2N}\Big(\sum_{x^o \in X^o}\sum_{i=1}^{l_{x^o}} \widehat{\mathcal{I}}^{JSD}(g(x^o); f_{\theta}^{(i)}(x^o))) + \sum_{x^a \in X^a}\sum_{i=1}^{l_{x^a}} \widehat{\mathcal{I}}^{JSD}(g(x^a); f_{\theta}^{(i)}(x^a))). \quad (6)$$

An estimated MI for each sequence $g(x)$ and token representations $f_{\theta}^{(i)}(x)$ is calculated as follows:

$$\widehat{\mathcal{I}}_{\theta}^{JSD}(g(x); f_{\theta}^{(i)}(x)) := E_{\mathbb{P}}[-sp(-g(x) \cdot f_{\theta}^{(i)}(x))] - E_{\mathbb{P}\times\tilde{\mathbb{P}}}[sp(g(x) \cdot f_{\theta}^{(i)}(\tilde{x}))], \quad (7)$$

where $\tilde{x}$ is a different text on the minibatch.

3.2 Clustering with KL Divergence

In addition to the MI objective, we employ $\mathcal{L}_{\text{Cluster}}$ during the learning stage to encourage the coalescence of samples that are most likely to belong to the same cluster. We follow the clustering method proposed by Xie et al. (2016), which is also used by Zhang et al. (2021). This method involves computing soft cluster assignments and formulating the clustering objective using KL divergence.

For the first step, we follow Xie et al. (2016) using the Student's t-distribution Q to compute a soft cluster assignment for each text instance $x_j \in X$ and the centroid μ_k where $\mu_k \in \{1, ..., K\}$ for the dataset with K-clusters. Specifically, we compute the probability q_{jk} of assigning a text x_j to a cluster μ_k as follows.

$$q_{jk} = \frac{(1 + \|g(x_j) - \mu_k\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K}(1 + \|g(x_j) - \mu_{k'}\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}} \quad (8)$$

The α symbol represents the degree of freedom of the distribution, and we set α to 1. Following Zhang et al. (2021), each centroid μ_k is approximated by the linear clustering head c_θ.

The second step is calculating an auxiliary target distribution P and using it to assist in refining clusters' centroids. The main idea is to give more importance to text samples with high clustering confidence. The probability $p_{jk} \in P$ is defined as

$$p_{jk} = \frac{q_{jk}^2 / \sum_{j'} q_{j'k}}{\sum_{k'}(q_{jk'}^2 / \sum_{j'} q_{j'k'})}. \quad (9)$$

To match the soft cluster assignments to the target distribution, the KL-divergence between probability distributions P and Q is computed as follows.

$$\ell_j^C = KL\left[p_j \| q_j\right] = \sum_{k=1}^{K} p_{jk} \log \frac{p_{jk}}{q_{jk}} \quad (10)$$

We then formulate it as a clustering objective for each minibatch of size N as

$$\mathcal{L}_{\text{Cluster}} = \sum_{j=1}^{N} \ell_j^C / N. \quad (11)$$

4 Experimental Setup

Datasets. Following previous works (Rakib et al., 2020; Zhang et al., 2021; Pugachev and Burtsev, 2021; Zheng et al., 2023), we conduct experiments and evaluate the performance of MIST on the eight standard short text clustering datasets. The descriptions of the datasets are provided in Appendix A.1

Implementation. We implement our model in PyTorch (Paszke et al., 2017) and use the *paraphrase-mpnet-base-v2* in Sentence Transformers library (Reimers and Gurevych, 2019b) as the encoder, with a linear clustering head following Zhang et al. (2021). The encoder is trained for 1,200 iterations for all datasets and we use Adam optimizer with a batch size of 256. The learning rate of the encoder and the clustering head are set to 6e−6 and 6e−5, respectively. We follow Xu et al. (2017) and Hadifar et al. (2019) by randomly selecting 10% of data as the validation set. Furthermore, we follow Zhang et al. (2021) by not performing any preprocessing operations on all eight datasets. Although some of the existing works preprocess the texts by removing symbols, stop words, and punctuation, or converting them to lowercase.

In the training stage, the original and augmented texts are taken into consideration as inputs for the MI objective $\mathcal{L}_{\text{MI}}$, since we found that they are more effective than employing two augmented

	AgNews		SearchSnippets		StackOverflow		Biomedical	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Reported in the References								
STCC	-	-	77.09	63.16	51.13	49.03	43.62	38.05
Self-Train	-	-	77.1	56.7	59.8	54.8	**54.8**	**47.1**
SCA-AE	68.36	34.14	68.71	50.26	76.55	65.99	40.25	33.29
HAC-SD	81.84	54.57	**82.69**	63.76	64.80	59.48	40.13	33.51
RSTC	84.24	62.45	80.10	69.74	**83.30**	74.11	48.40	40.12
Reimplementation								
SBERT (k-means)	83.44	57.76	73.02	59.77	76.79	75.12	41.30	36.93
SCCL	85.67	65.98	78.73	70.10	78.35	75.6	39.35	39.2
SCCL-Multi	85.6	66	78.6	**70.17**	78.3	76.22	39.2	33.7
Proposed Method								
MIST	**89.47***	**70.25***	76.72	67.69	79.65	**78.59***	39.15	34.66

	Tweet		GoogleNews-TS		GoogleNews-T		GoogleNews-S	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Reported in the references								
STCC	-	-	-	-	-	-	-	-
Self-Train	-	-	-	-	-	-	-	-
SCA-AE	84.85	89.19	-	-	-	-	-	-
HAC-SD	89.62	85.20	85.76	88.00	**81.75**	84.20	80.63	83.50
RSTC	75.20	87.35	83.27	93.15	72.27	87.39	79.32	89.40
Reimplementation								
SBERT (k-means)	62.7	86.8	67.40	90.47	63.98	86.13	65.87	87.64
SCCL	68.3	88.59	78.9	92.92	69.9	87.9	73.55	89.33
SCCL-Multi	67.55	88.41	80.15	93.4	72.85	88.44	74.2	89.47
Proposed Method								
MIST	**91.75***	**95.12***	**90.63***	**96.42***	78.8	**89.31***	**82.14***	**90.86***

Table 1: Experimental results on eight short text clustering datasets. * denotes that MIST is significantly better than both reimplemented versions of SCCL. In order to statistically compare models, we use the Almost Stochastic Dominance test (Dror et al., 2019) with the significant level of 0.05.

pairs. We follow Zhang et al. (2021) and utilize *Contextual Augmenter* (Kobayashi, 2018; Ma, 2019) to generate augmented samples for each text instance as it was demonstrated to produce the best outcomes in their study. To assess clustering performance, we use the same standard metrics—Accuracy (ACC) and Normalized Mutual Information (NMI)— as used in all competitive methods [1]. The results are averaged over five trials.

5 Experimental Results

We extensively compare the performance of MIST with state-of-the-art methods including STCC (Xu et al., 2017), Self-Train (Hadifar et al., 2019), HAC-SD (Rakib et al., 2020), SCA-AE (Yin et al., 2021), SCCL (Zhang et al., 2021), and RSTC (Zheng et al., 2023). In addition, this section provides an ablation study on our proposed method.

5.1 Main Results

As shown in Table 1, MIST achieves state-of-the-art results in terms of Accuracy and NMI for most cases on the eight benchmark datasets. In contrast, HAC-SD and Self-Train attain the best results in only two cases, whereas SCCL and RSTC produce the best outcome in only one case. Note that, the performances of MIST are collected using the identical setting and training iteration across all datasets to demonstrate generalizability. As a result, the need for a specific configuration for each dataset is avoided, enabling a reduction in model overhead.

For datasets with a small number of clusters in the upper section of the table, MIST shows superior performances on AgNews for both metrics and StackOverflow in terms of NMI. Notably, there are two datasets that MIST is outperformed by competitors for both ACC and NMI, i.e., Biomedical and SearchSnippets. For Biomedical, Hadifar et al. (2019) dominates the competitive methods. They achieve the best results by using an in-domain pretrained model to process this dataset, whereas the dataset used to pretrain our encoder and other recent methods is a general-domain one.

For SearchSnippets, we observe that most of the text samples are collections of keywords and terminologies rather than coherent sentence structures. Moreover, SearchSnippets samples are medium-

[1]The Accuracy is calculated via the Hungarian algorithm, and NMI measures the information shared between the ground truth assignments and the predicted assignments.

length fragmental sequences; as a result, the token-level MI maximization objective is more emphasized due to the length of the texts. These two factors exert a direct impact on the token-level MI maximization objective while it is being executed in the learning stage. Since the token vectors are contextualized representations, forcing the model to learn from incoherent contextual signals can be detrimental to the overall sequence representations, which are subsequently used in the clustering stage. This can be more problematic when the same keywords appear in multiple clusters.

As demonstrated in the lower section of the table, MIST obtains the best outcomes on most of the datasets containing a large number of clusters. Due to the fine-grained categorization of these datasets, texts in different clusters may share similar content or keywords, hence inducing ambiguity. This ambiguity in textual data and ground truths leads to erroneous predictions. Moreover, another cause of inaccuracy is when the text content in one cluster is a subtopic of another. GoogleNews-T, which only contains news headlines that are relatively short with few keywords, presents an additional challenge for clustering these extremely short texts into a large number of clusters. In terms of Accuracy, our method achieves a result comparable to that of Rakib et al. (2020) on GoogleNews-T. We conjecture that hierarchical clustering and outlier removal algorithms employed in their method can better deal with the hierarchical nature of data in this scenario. However, MIST outperforms Rakib et al. (2020) in terms of NMI on this dataset.

Although GoogleNews-S and GoogleNews-TS share the same challenges as GoogleNews-T, clustering texts in both datasets is more accurate due to the benefit of additional context and information in the texts themselves. As GoogleNews-S contains snippets of news, and GoogleNews-TS includes both titles and snippets. Consequently, MIST achieves superior clustering performances on both datasets for both matrices.

Furthermore, we thoroughly compare MIST with SCCL, as this current state-of-the-art model also utilizes contrastive learning and aims to improve the effectiveness of representations for short text clustering, which is similar to our contribution, by reproducing SCCL in two versions for a fair comparison: an end-to-end (original) version, and a multi-stage version. For the latter, we apply the k-means algorithm on the trained representations to get the final clusters, referred to as *SCCL-Multi*. In particular, SCCL-Multi is analogous to our framework, except for the representation learning technique. The reimplemented versions use the same backbone and augmentation setting as our model.

The comparative results show that MIST outperforms SCCL for both versions in most cases. More specifically, the superior performances of MIST compared to SCCL-Multi demonstrate that our proposed representation learning procedure improves short text representations more effectively than the standard contrastive learning objective in the SCCL framework. MIST also consistently surpasses both reimplemented versions of SCCL in other settings, including settings indicated in their publication in most cases, as shown in Appendix A.6.

5.2 Ablation Study

To better understand the effect of the various model modifications on the clustering performance and the analysis versus text lengths, we conducted additional experiments by varying the trade-off between components in our training procedure.

5.2.1 The Impact of Sequence- and Token-MI Maximization Objectives

This experiment studies the impact of the ratio between two MI maximization objectives on the clustering performance and the importance of incorporating both objectives in our representation learning procedure. We report and analyze the performance of our model using four different values of λ. Particularly, λ denotes the weight of token-level MI maximization objectives, $\mathcal{I}_{\text{tok}}$, and 1-λ represents the weight of sequence-level MI objectives, $\mathcal{I}_{\text{seq}}$. We consider the following settings: (1) *MIST-seq*: our model with a sequence-only MI maximization objective ($\lambda = 0$), (2) *MIST-tok*: our model with a token-only MI maximization objective ($\lambda = 1$), (3) *MIST-equal*: our model with both objectives are given an equivalent weight ($\lambda = 0.5$), and (4) *MIST*: our proposed version, i.e., our model with the λ determined by the adaptive weighting function, Eq.3, varying according to input text length.

As shown in Figure 2, MIST with the value of λ set by Eq.3 yields the best performances in terms of Accuracy for most datasets and shows performance gains compared to other settings. We also discovered that NMI tends to follow the same trend as Accuracy, as presented in Appendix A.2. This demonstrates that the length of short texts has a great impact on determining the appropriate ratio

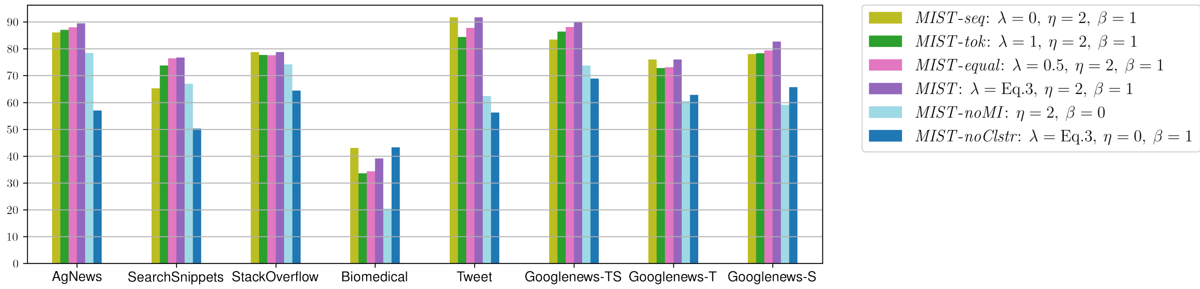

Figure 2: Accuracy for six different settings including four different weighting ratios between sequence-level and token-level MI maximization objectives. As well as, a setting where the clustering objective is absent ($\eta = 0$), and a setting where the MI objective is absent ($\beta = 0$). Note that when we set β to 0, λ has no effect.

between the two MI objectives, i.e. the optimal ratio varies by input samples. By utilizing the proposed adaptive weighting function, MIST can perform effectively across various datasets.

For medium or large fragmental sequences, such as GoogleNews-TS, MIST produces the best outcomes when the weight λ calculated by Eq.3—the value of λ is greater than 0. Remarkably, MIST-equal and MIST-tok always outperform MIST-seq in this situation. This shows that only the sequence-level objective is inadequate when dealing with lengthy texts, as larger fragments usually have a higher signal-to-noise ratio. However, this issue can be mitigated by performing the token-level MI maximization during the learning stage.

For small fragment datasets, such as Tweet, text samples are relatively short and contain less signal-to-noise problem. In this scenario, the weight λ is equal to 0 based on Eq.3, i.e., MIST is identical to MIST-seq, which outperforms all other settings. MIST-tok and MIST-equal may encourage the encoder to learn text representations by placing emphasis on keywords that could also appear in multiple clusters, causing ambiguity and error in clustering. Hence, the token MI objective provides advantages when used in a suitable weight.

In addition, we investigate the situation in which the MI objective is removed ($\beta = 0$), *MIST-noMI*. The ablation results show significant drops in the performance on all datasets. This implies that the MI objective is essential for performance gain.

5.2.2 The Impact of the Clustering Objective

As shown in Figure 2, the clustering performance drops drastically when we remove the clustering objective ($\eta = 0$) during learning representations, *MIST-noClstr*. This demonstrates that the categorical structure imposed by jointly optimizing the clustering objective with the MI objective is a crucial component that boosts performance. Furthermore, we observe that as the weight of the clustering objective (η) increases, the performances continuously improve until η reaches its saturation point at 2. In Figure 3, the average Accuracy and NMI for all eight datasets improve as the clustering weight is steadily increased until it reaches 2.

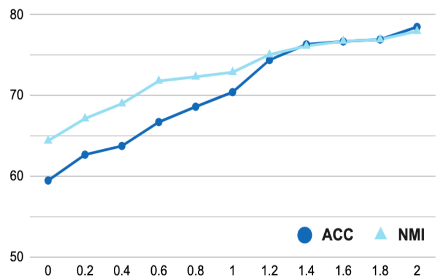

Figure 3: The average clustering performance across eight datasets based on the clustering objective strength.

6 Conclusion

We propose a novel multi-stage short text clustering framework that mainly focuses on improving the representation learning stage. Our adaptive learning approach integrates two MI maximization objectives operating at the sequence and token levels to produce effective representations. This mechanism allows us to simultaneously learn distinct text representations while maintaining limited information in a weak signal environment. In addition, we introduce a generalized adaptive weighting function that considers the length of the texts to determine an optimal ratio between the two MI maximization objectives during the learning stage.

MIST outperforms competitive methods in most cases in terms of Accuracy and NMI across eight benchmark datasets. This demonstrates that utilizing the MI maximization strategy for learning representation in a constrained environment could potentially be a promising tactic. Further study would be worthwhile since it might enhance the quality of textual representations for other tasks.

Limitations

This section discusses the limitations of our proposed framework. Firstly, the encoder of our model is pretrained using general domain data. Hence, the performance drops when our model encounters short texts in a specific domain, such as Biomedical. Furthermore, short text inputs containing only of keywords or incoherent text sequences hinder the performance of our representation learning method. In particular, when dealing with lengthy texts that lack coherence, optimizing both token-level MI and sequence-level MI maximization forces a sequence representation to resemble each individual token embedding. The token-level MI maximization objective provides no further improvement in this case. This issue is exacerbated when some terms are shared across clusters. This constraint should be taken into account in future research.

Another limitation involving the general operation of contrastive learning is that the selection of augmented samples directly affects the clustering performance. Notably, the best augmentation strategy is still a subject of discussion and needs more exploration. A study in Zhang et al. (2021) and our own experiments with various augmentation settings show that varying an augmenter as well as adjusting the configuration parameters both affect the performance. Additionally, even if the augmenter and the parameters used to generate augmented texts are exactly the same, there is a possibility that the outcomes from the two trials may vary, adding a variance to the performance results.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15509–15519.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy. Association for Computational Linguistics.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 879–895. Association for Computational Linguistics.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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)*. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam

Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 452–457. Association for Computational Linguistics.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-aware contrastive learning of sentence embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3870–3885, Seattle, United States. Association for Computational Linguistics.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 271–279.

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 2017 Workshop on Autodiff*.

Xuan Hieu Phan, Minh Le 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, WWW 2008, Beijing, China, April 21-25, 2008*, pages 91–100. ACM.

Leonid Pugachev and Mikhail S. Burtsev. 2021. Short text clustering with transformers. *CoRR*, abs/2102.00541.

Md. Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, and Evangelos E. 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, NLDB 2020, Saarbrücken, Germany, June 24-26, 2020, Proceedings*, volume 12089 of *Lecture Notes in Computer Science*, pages 105–117. Springer.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Nils Reimers and Iryna Gurevych. 2019b. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoffrey J. Gordon, Philip Bachman, and Remi Tachet des Combes. 2021. Decomposed mutual information estimation for contrastive representation learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9859–9869. PMLR.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 478–487. JMLR.org.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, pages 62–69. The Association for Computational Linguistics.

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

Hui Yin, Xiangyu Song, Shuiqiao Yang, Guangyan Huang, and Jianxin Li. 2021. Representation learning for short text clustering. In *Web Information Systems Engineering - WISE 2021 - 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26-29, 2021, Proceedings, Part II*, volume 13081 of

Lecture Notes in Computer Science, pages 321–335. Springer.

Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, pages 625–636. IEEE Computer Society.

Dessalew Yohannes and Yeregal Assabie. 2021. Amharic text clustering using encyclopedic knowledge with neural word embedding. *CoRR*, abs/2105.00809.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen R. McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. *CoRR*, abs/2103.12953.

Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew Arnold. 2022. Virtual augmentation supported contrastive learning of sentence representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 864–876, Dublin, Ireland. Association for Computational Linguistics.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *CoRR*, abs/1502.01710.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1601–1610. Association for Computational Linguistics.

Xiaolin Zheng, Mengling Hu, Weiming Liu, Chaochao Chen, and Xinting Liao. 2023. Robust representation learning with reliable pseudo-labels generation via self-adaptive optimal transport for short text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10493–10507, Toronto, Canada. Association for Computational Linguistics.

A Appendices

A.1 Datasets

Following previous works (Rakib et al., 2020; Zhang et al., 2021; Pugachev and Burtsev, 2021; Zheng et al., 2023), we conduct experiments and assess the performance of our model on eight English benchmark datasets for short text clustering. Table 2 presents the important statistics of all datasets.

- **AgNews**: a subset of the English news titles dataset (Zhang and LeCun, 2015) in 4 different topics, with 2,000 samples chosen randomly from each topic by Rakib et al. (2020).
- **SearchSnippets**: a dataset consisting of 12,340 web search snippets from 8 different categories (Phan et al., 2008).
- **Biomedical**: 20,000 paper titles, from 20 different Medical Subject Headings (MeSH), randomly selected by Xu et al. (2017) from the PubMed data distributed by BioASQ3.
- **StackOverflow**: challenge data published on Kaggle and randomly chosen by Xu et al. (2017), comprising 20,000 questions from Stack Overflow related to 20 distinct tags.
- **Tweet**: a dataset comprising 2,472 tweets with 89 groups (Yin and Wang, 2016).
- **GoogleNews**: GoogleNews-TS is a collection of titles and text snippets from 11,109 news articles covering 152 events (Yin and Wang, 2016). Only titles and snippet of each news article were extracted to produce GoogleNews-T and GoogleNews-S, respectively.

Dataset	$N^{Cluster}$	N^{Doc}	N^{Word}
AgNews	4	8,000	23
SearchSnippets	8	12,340	18
Biomedical	20	20,000	13
StackOverflow	20	20,000	8
Tweet	89	2,472	8
Googlenews-TS	152	11,109	28
Googlenews-T	152	11,109	6
Googlenews-S	152	11,109	22

Table 2: Dataset statistics. $N^{Cluster}$: number of clusters; N^{Doc}: number of short text documents; N^{Word} : average number of words in each document

We spend up to 14 GPU hours on a Tesla V100 32G GPU to complete the training on all datasets for each MIST model's configuration.

A.2 The Effects of Sequence- and Token-MI Maximization Objectives on NMI

Figure 4 shows the impact of the different ratios between the two MI maximization objectives on the clustering performance in terms of NMI across eight short text datasets. It follows the same trend as Accuracy as discussed in Section 5.2.1. MIST with our proposed generalized adaptive weighting function obtains the best clustering performance in terms of NMI for most datasets.

A.3 Positive Pairs in Contrastive Learning

It is a common practice in contrastive learning frameworks to only consider augmented texts as inputs, excluding original samples. However, we

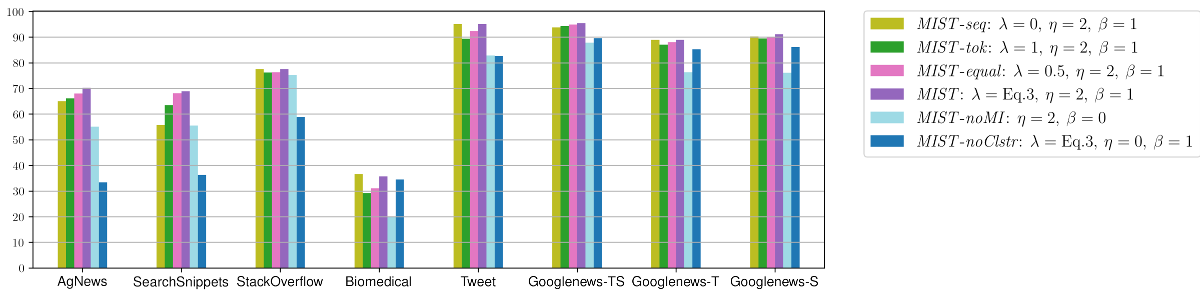

Figure 4: NMI for six different settings including four different weighting ratios between sequence-level and token-level MI maximization objectives. As well as, a setting where a clustering loss is absent ($\eta = 0$), and a setting where an MI loss is absent ($\beta = 0$). Note that when we set β to 0, λ has no effect.

	AgNews		**SearchSnippets**		**StackOverflow**		**Biomedical**	
	ACC	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
MIST w/ $\eta = 0$	56.96	33.40	50.30	36.30	64.40	58.80	43.26	34.55
MIST w/ $\eta = 1$	81.40	57.39	70.99	56.90	76.41	71.92	47.66	40.34
MIST w/ $\eta = 2$	89.47	70.25	76.72	67.69	78.74	77.59	39.15	34.66

	Tweet		**GoogleNewsTS**		**GoogleNewsT**		**GoogleNewsS**	
	ACC	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
MIST w/ $\eta = 0$	56.27	82.64	68.89	89.59	62.85	85.28	65.74	86.16
MIST w/ $\eta = 1$	64.46	86.27	74.86	91.89	66.91	87.04	71.98	88.58
MIST w/ $\eta = 2$	91.75	95.12	89.93	95.47	75.97	88.97	81.91	90.79

Table 3: The clustering results of MIST on three different weights of the clustering objective, η.

adopt a different input scheme. We discovered that feeding both original and augmented samples into our representation learning framework (as shown in Figure 1) yields better clustering results than exclusively taking two augmented texts as an input pair. One plausible reason is that when augmented texts are generated, the augmenter replaces some keywords in the original texts with new words. Short texts inherently have few keywords; hence, the absence of crucial words required for text categorization impacts clustering performance.

A.4 The Analysis of the Clustering Objective

As discussed in Section 5.2.2, the clustering performance is substantially affected by the weight of the clustering objective. Table 3 presents the performance of MIST across eight datasets in three situations, i.e., the coefficient of the clustering objective, η, in Eq.1 is assigned to 0, 1, and 2. The optimal results for the majority in terms of ACC and NMI are produced when η is set to 2.

A.5 Exploration of Data Augmentations

According to Zhang et al. (2021), which has studied the impacts of data augmentation in extensive details. The *Contextual Augmenter* has shown that it substantially outperforms other augmenters in their study. They hypothesized that since both the Contextual Augmenter and their encoder use the pretrained transformers as the backbones, this allows the Contextual Augmenter to produce augmentation texts that are more informative and beneficial to their framework. We also adopted a pretrained transformer as the encoder in our framework and we observed that the experimental results followed the same trend as Zhang et al. (2021). We thus employ this augmenter in our experiments.

In this section, we investigate the impact of the Contextual Augmenter configurations in terms of masked language models and word substitution ratios. As shown in Table 4, we found that MIST using augmented texts generated from the BERT model with *20% substitution rate* yields the best overall performance. Interestingly, MIST with augmented texts produced by *other masked language models with a 20% substitution rate* also yields outcomes close to those of BERT with the same substitution rate.

A.6 SCCL Reimplementation

To thoroughly compare the performance of our proposed representation learning strategy against the standard contrastive learning method in SCCL (Zhang et al., 2021), we reproduced SCCL in both an end-to-end version (SCCL) and a multiple-stage

	AgNews		SearchSnippets		StackOverflow		Biomedical	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
MIST w/ BERT 10%	87.74	66.99	75.98	67.71	77.78	76.42	37.51	33.97
MIST w/ BERT 20%	89.47	70.25	76.72	67.69	78.74	77.59	39.15	34.66
MIST w/ BERT 30%	86.33	66.09	81.46	67.71	73.60	71.55	39.79	34.61
MIST w/ RoBERTa 10%	87.51	66.81	75.64	67.11	77.84	76.50	38.61	35.11
MIST w/ RoBERTa 20%	88.85	69.12	76.21	68.52	77.74	76.41	37.17	31.62
MIST w/ RoBERTa 30%	86.43	66.4	73.77	65.72	77.76	77.03	29.48	27.38
MIST w/ DistilBERT 10%	87.22	66.44	74.96	65.89	77.67	76.30	38.29	34.29
MIST w/ DistilBERT 20%	89.42	70.26	75.74	67.85	77.72	77.05	38.29	32.31
MIST w/ DistilBERT 30%	87.96	67.66	74.23	64.11	77.67	76.34	38.83	34.63

	Tweet		GoogleNews-TS		GoogleNews-T		GoogleNews-S	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
MIST w/ BERT 10%	88.76	93.04	86.65	94.76	72.41	87.99	76.56	89.3
MIST w/ BERT 20%	91.75	95.12	89.93	95.47	75.97	88.97	81.91	90.79
MIST w/ BERT 30%	90.07	94.14	89.28	94.98	75.63	88.55	80.74	89.99
MIST w/ RoBERTa 10%	88.18	92.64	85.85	94.48	73.68	88.00	77.89	89.52
MIST w/ RoBERTa 20%	90.97	94.67	90.10	95.35	74.61	88.27	77.62	90.00
MIST w/ RoBERTa 30%	83.40	95.15	88.29	96.20	70.27	88.24	78.43	89.82
MIST w/ DistillBERT 10%	85.48	92.24	85.15	94.42	75.89	88.51	77.55	89.69
MIST w/ DistillBERT 20%	91.24	94.99	90.16	95.43	74.14	88.53	82.54	90.69
MIST w/ DistillBERT 30%	86.56	92.50	85.85	94.46	75.57	88.50	77.18	89.52

Table 4: The clustering performance of MIST when feeding augmented texts generated by Contextual Augmenter as inputs across nine different configurations.

version (SCCL-Multi). For the latter version, we apply the k-means algorithm on top of SCCL representations to make their pipeline identical to our framework except for the representation learning method. To be more specific, in this study, we report the experimental results of both reimplemented versions of SCCL using the *backbone* identified in the experimental setup of their publication. Moreover, SCCL considers the *Contextual Augmenter* with three configurations by setting the *word substitution ratio* of each text instance to 10%, 20%, and 30%. However, their study does not identify which setting produces the best outcomes. Therefore, we evaluate both reproduced versions of SCCL using *three alternative masked language models*: BERT-base, RoBERTa, and DistilBERT, with the aforementioned word substitution ratios for augmented pair generation to cover all scenarios reported in their study.

Table 5 reports the clustering performance of SCCL in both reproduced versions and in all configurations mentioned above. The reported performances show that despite the reproduced SCCL employing the configuration specified in their reference paper, their outcomes are still inferior to MIST in most cases. More specifically, MIST with the setup described in Section 4 outperforms SCCL and SCCL-Multi with the best parameter settings in the majority of cases. The fact that MIST produces better clustering performance than SCCL-Multi in this study emphasizes that our proposed representation learning technique improves short text representations more effectively than the standard contrastive learning objective in the SCCL framework for short text clustering task. This demonstrates the success and efficiency of our proposed learning method even when compared with SCCL in various settings. Note that we collected the experimental results of reimplemented versions of SCCL from the *best iteration* for each dataset throughout 3000 iterations instead of using a stopping criterion, which is not indicated in their publication. Besides, the performances in their publication are reported from multiple settings.

Interestingly, the percentage of word replacement and masked language models employed for augmented text generation have an impact on the clustering performance. The best setting for these two parameters varies across different datasets. However, the performances of our proposed method presented in Table 1 are reported by using only a single setting for all datasets.

	AgNews		SearchSnippets		StackOverflow		Biomedical	
	ACC	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
SCCL (in the reference paper)	88.20	68.20	85.20	71.10	75.50	74.50	46.20	41.50
SCCL w/ BERT 10%	87.20	66.94	83.70	70.05	71.40	71.28	46.00	40.06
SCCL-Multi w/ BERT 10%	87.2	66.94	83.40	69.88	77.30	73.76	46.00	40.13
SCCL w/ BERT 20%	87.10	66.91	84.40	69.58	64.20	56.23	46.40	40.39
SCCL-Multi w/ BERT 20%	87.10	66.80	83.60	69.28	60.02	52.22	45.50	40.07
SCCL w/ BERT 30%	87.50	67.46	83.70	68.54	60.70	52.18	42.40	38.14
SCCL-Multi w/ BERT 30%	87.50	67.45	82.60	66.45	60.90	52.29	42.30	37.95
SCCL w/ RoBERTa 10%	87.00	66.57	84.50	70.21	62.10	54.26	28.50	20.35
SCCL-Multi w/ RoBERTa 10%	87.00	66.55	84.10	70.14	61.40	53.05	28.50	20.34
SCCL w/ RoBERTa 20%	85.20	64.20	62.60	41.66	60.70	52.26	39.60	32.66
SCCL-Multi w/ RoBERTa 20%	85.10	64.24	72.00	51.23	60.09	52.31	38.40	38.40
SCCL w/ RoBERTa 30%	84.00	62.24	30.70	10.07	60.70	52.28	39.10	32.77
SCCL-Multi w/ RoBERTa 30%	84.00	62.26	30.70	10.05	60.90	52.44	39.50	32.63
SCCL w/ DistilBERT 10%	87.30	67.16	84.70	70.79	70.20	69.49	46.10	39.87
SCCL-Multi w/ DistilBERT 10%	87.30	67.16	84.50	70.64	72.10	68.20	46.20	39.92
SCCL w/ DistilBERT 20%	86.80	65.87	84.70	70.62	71.40	69.38	46.30	39.94
SCCL-Multi w/ DistilBERT 20%	86.80	65.87	84.20	70.45	72.20	70.84	46.40	40.01
SCCL w/ DistilBERT 30%	87.20	66.77	85.00	71.63	70.80	70.04	46.30	40.49
SCCL-Multi w/ DistilBERT 30%	87.20	66.75	84.60	71.35	76.50	72.57	46.40	40.58

	Tweet		GoogleNews-TS		GoogleNews-T		GoogleNews-S	
	ACC	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**	**ACC**	**NMI**
SCCL (in the reference paper)	78.20	89.20	89.80	94.90	75.80	88.30	83.10	90.40
SCCL w/ BERT 10%	56.80	81.91	70.10	89.49	62.50	81.53	69.00	86.29
SCCL-Multi w/ BERT 10%	75.30	88.39	86.70	93.95	76.30	88.25	81.00	89.82
SCCL w/ BERT 20%	57.10	82.54	75.60	90.99	63.00	81.72	67.80	85.97
SCCL-Multi w/ BERT 20%	78.20	89.41	88.70	94.70	76.20	87.97	81.10	89.60
SCCL w/ BERT 30%	56.6	82.23	74.2	90.83	61.30	81.20	64.9	89.78
SCCL-Multi w/ BERT 30%	78.80	89.58	89.90	94.91	75.60	87.88	82.10	89.77
SCCL w/ RoBERTa 10%	56.00	79.89	73.60	90.46	55.60	78.08	65.50	85.26
SCCL-Multi w/ RoBERTa 10%	71.10	85.86	86.60	93.94	56.90	78.52	80.50	89.50
SCCL w/ RoBERTa 20%	56.80	79.56	74.90	90.37	55.60	78.08	66.90	85.38
SCCL-Multi w/ RoBERTa 20%	74.20	86.61	88.10	94.27	58.40	79.28	81.30	89.87
SCCL w/ RoBERTa 30%	53.80	78.47	71.80	71.80	55.60	78.42	65.30	83.99
SCCL-Multi w/ RoBERTa 30%	63.60	76.98	85.20	93.53	56.60	78.42	78.00	88.14
SCCL w/ DistilBERT 10%	56.10	80.87	72.70	90.03	61.40	80.94	69.60	85.81
SCCL-Multi w/ DistilBERT 10%	78.80	88.91	87.70	94.25	74.30	87.78	79.70	89.20
SCCL w/ DistilBERT 20%	56.40	80.28	71.70	90.04	61.30	81.19	67.70	86.02
SCCL-Multi w/ DistilBERT 20%	77.10	88.61	86.50	94.03	75.10	87.51	79.50	89.70
SCCL w/ DistilBERT 30%	56.60	81.65	72.10	90.18	62.00	81.09	66.50	85.48
SCCL-Multi w/ DistilBERT 30%	76.00	88.39	88.50	94.18	75.80	87.60	79.10	89.01

Table 5: The clustering performances of the reimplemented SCCL and SCCL-Multi with nine different configurations for Contextual Augmenter. These configurations are obtained by setting the word substitution ratio of each text instance to 10% , 20%, and 30%, as well as using three alternative masked language models: BERT-base, RoBERTa, and DistilBERT.