Contrastive Learning Models for Sentence Representations

LINGLING XU, Hong Kong Metropolitan University, Hong Kong SAR
HAORAN XIE, Lingnan University, Hong Kong SAR
ZONGXI LI, FU LEE WANG, and WEIMING WANG, Hong Kong Metropolitan University, Hong Kong SAR
QING LI, The Hong Kong Polytechnic University, Hong Kong SAR

Sentence representation learning is a crucial task in natural language processing, as the quality of learned representations directly influences downstream tasks, such as sentence classification and sentiment analysis. Transformer-based pretrained language models such as bidirectional encoder representations from transformers (BERT) have been extensively applied to various natural language processing tasks, and have exhibited moderately good performance. However, the anisotropy of the learned embedding space prevents BERT sentence embeddings from achieving good results in the semantic textual similarity tasks. It has been shown that contrastive learning can alleviate the anisotropy problem and significantly improve sentence representation performance. Therefore, there has been a surge in the development of models that utilize contrastive learning to fine-tune BERT-like pretrained language models to learn sentence representations. But no systematic review of contrastive learning models for sentence representations has been conducted. To fill this gap, this article summarizes and categorizes the contrastive learning based sentence representation models, common evaluation tasks for assessing the quality of learned representations, and future research directions. Furthermore, we select several representative models for exhaustive experiments to illustrate the quantitative improvement of various strategies on sentence representations.

CCS Concepts: • **General and reference** → **Surveys and overviews** • **Computing methodologies** → **Natural language processing**;

Additional Key Words and Phrases: Sentence representation learning, contrastive learning, Data Augmentation, BERT

ACM Reference format:
Lingling Xu, Haoran Xie, Zongxi Li, Fu Lee Wang, Weiming Wang, and Qing Li. 2023. Contrastive Learning Models for Sentence Representations. *ACM Trans. Intell. Syst. Technol.* 14, 4, Article 67 (June 2023), 34 pages. https://doi.org/10.1145/3593590

This research was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19) and the Lam Woo Research Fund (LWP20019) of Lingnan University.
Authors' addresses: L. Xu, Z. Li, F. L. Wang, and W. Wang, Hong Kong Metropolitan University, 30 Good Shepherd Street, Ho Man Tin, Kowloon, Hong Kong SAR; emails: xxiao199409@gmail.com, zongxili2@gmail.com, pwang@hkmu.edu.hk, wmwang@hkmu.edu.hk; H. Xie (corresponding author), Lingnan University, 8 Castle Peak Road, Tuen Mun, NewTerritories, Hong Kong SAR; email: hrxie2@gmail.com; Q. Li, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong SAR; email: qing-prof.li@polyu.edu.hk.

2157-6904/2023/06-ART67
https://doi.org/10.1145/3593590

1 INTRODUCTION

Learning sentence representations[1] has long been an important research hotspot in **Natural Language Processing (NLP)**. Sentence representations encapsulate key semantic and syntactic information about sentences, which directly affects the performance of downstream tasks, such as sentence classification [60, 61], sentiment analysis [99, 135], and semantic matching [37, 66]. Numerous sentence representation learning models have been proposed, among which BERT (the bidirectional encoder representations from transformers) [27], a transformer-based pretrained model, stands out for its excellent performance in various NLP tasks.

However, some studies [29, 34, 58] have found that the sentence representation performance of transformer-based **Pretrained Language Models (PLMs)** is constrained by the anisotropic embedding space, in which word embeddings occupy a narrow cone in the vector space. In other words, high-frequency words are close to the origin and closely dispersed, whereas low-frequency words are far away from the origin and sparsely dispersed. Therefore, even if a high-frequency word and a low-frequency word are semantically equivalent, the difference in word frequency produces a large distance bias such that the distance between word embeddings cannot accurately indicate their semantic similarity. This results in the BERT sentence representations failing to outperform the non-contextualized average GloVe embeddings [84] in the **Semantic Textual Similarity (STS)** tasks, as indicated by Li et al. [58] and Reimers and Gurevych [88]. BERT-flow [58] and BERT-whitening [98] have been proposed to address the anisotropy problem by converting the BERT embedding space into an isotropic space.

In the work of Gao et al. [35], researchers theoretically and empirically demonstrated that the contrastive learning based sentence representation model SimCSE can ease the anisotropy problem by pushing negative pairs apart, and optimize alignment by pulling positive pairs close, which cannot be achieved in BERT-flow and BERT-whitening. Positive pairs are usually different views of the same instance, which are generated through various **Data Augmentation (DA)** strategies. Negative pairs are typically the remaining in-batch samples. With the success of SimCSE, fine-tuning BERT-like PLMs [27, 63, 90] through contrastive learning has become an increasingly popular strategy for generating high-quality sentence representations. These contrastive learning based sentence embedding models, such as SimCSE [35], MixCSE [138], and DiffCSE [22], yield competitive performance without supervision, even outperforming the state-of-the-art supervised sentence embedding model sentence-BERT (SBERT) [88] in STS tasks.

The two dominant frameworks,[2] SimCLR [16] and MoCo [41], have greatly facilitated the development of contrastive learning for their excellent performance in unsupervised representation learning. In addition, the contrastive learning performance is further enhanced by effective positive and negative samples. Section 3 focuses on the classification of sentence representation models using contrastive learning. Numerous studies [12, 35, 36, 48, 100, 119, 125, 137] have attempted to design various DA strategies to generate positive pairs, whereas others [59, 133, 134, 138, 142] have centered on generating useful negative samples to enhance sentence representation learning. Furthermore, some studies [22, 76, 107, 136] have demonstrated that incorporating external training data can help capture additional semantic features.

[1]We may interchangeably use the terms *representation* and *embedding* in this article.

[2]We notice that there are some disputes about whether to classify unsupervised representation learning models that do not involve negative samples (e.g., BYOL (Bootstrap Your Own Latent), SimSiam, and Barlow Twins, inter alia) as contrastive learning models. Hereby, we specify that only models involving the comparison of positive and negative pairs are strictly defined as contrastive learning models. Therefore, BYOL, SimSiam, and BarlowTwins, among others, are not within the scope of our review.

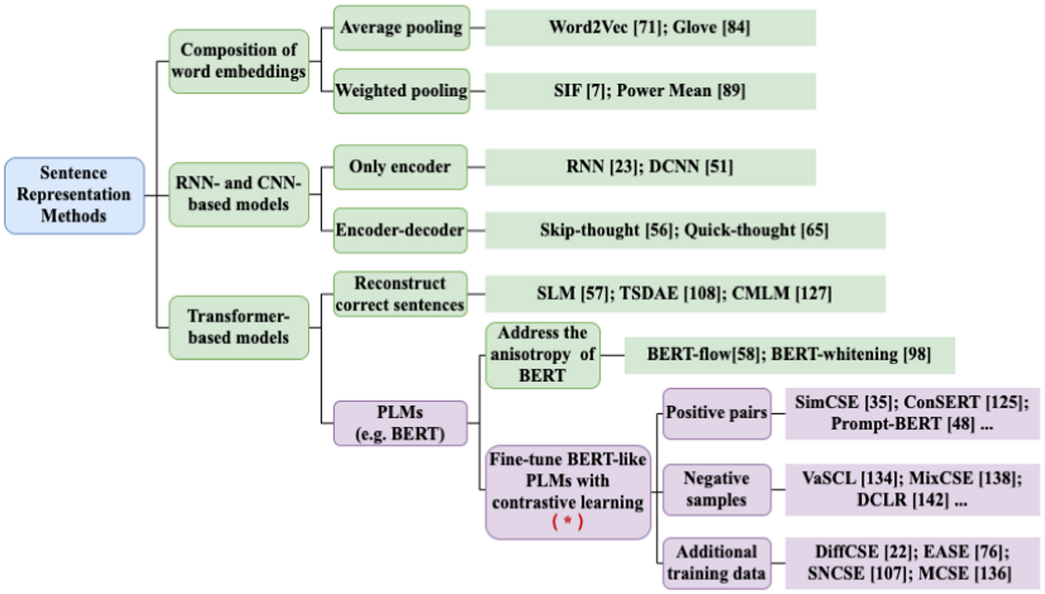

Fig. 1. The focus of this article (with *).

To assess the quality of the learned sentence representations, we discuss common evaluation tasks in Section 4, mainly including the STS tasks [1, 2, 3, 4, 5, 13, 68], transfer tasks [25], and short text clustering tasks [85, 86, 121, 129]. Section 5 provides a quantitative study of several representative models and their performance on the aforementioned evaluation tasks. Compared with vanilla BERT, nearly all of the contrastive learning models exhibited significant improvements on STS tasks and short text clustering tasks in terms of Spearman's correlation and clustering accuracy. We also analyzed the alignment and uniformity [109] performance of these models on the development set of the STS benchmark [13]. Moreover, the t-SNE [102] visualization of Stack Overflow [121] representations derived by these models showed that high-quality representations could better group semantically similar sentences. Finally, we outline potential research directions.

Our article aims to provide a comprehensive and systematic review of contrastive learning based sentence representation models shown in Figure 1. Given the extensive studies in this domain, we identify an appropriate method for categorizing recently proposed models and present the evolution of the related studies. This review provides a bird's eye view of the development trends, the limitations associated with existing methods, and future research directions in this field. To the best of our knowledge, this is the first in-depth survey on contrastive learning based sentence representation approaches. The main contributions to this survey are as follows:

- We identify strategies for enhancing the performance of contrastive sentence representation learning from the perspective of positive pairs, negative samples, and external training data.
- We summarize various evaluation tasks and datasets involved in sentence representation learning.
- We conduct exhaustive quantitative experiments to quantify and compare the effects of various improvement strategies on sentence representation learning and also demonstrate the effectiveness of contrastive learning.

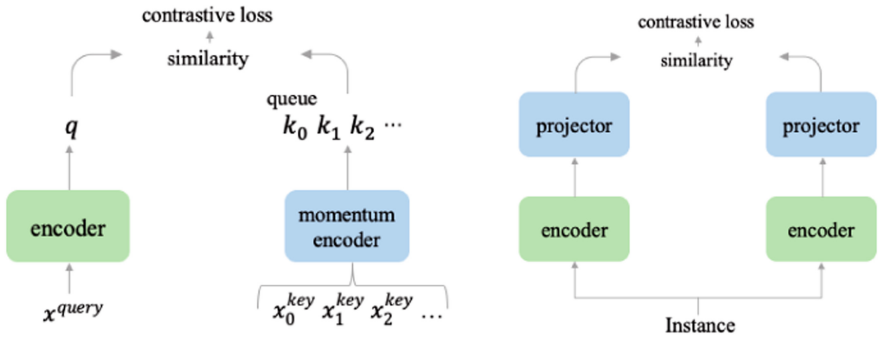

Fig. 2. Framework of MoCo (left) [41] and SimCLR (right) [16].

2 BACKGROUND

2.1 Contrastive Learning

This survey focuses on sentence representation models based on contrastive learning. First, we review contrastive learning to highlight the role of the contrastive learning algorithm. Contrastive learning is defined as a learning scenario in which the model is trained to pull semantically similar (positive) pairs together and push dissimilar (negative) pairs apart [40]. The outstanding performance of contrastive learning in the unsupervised setting has increased its popularity.

2.1.1 Contrastive Learning Framework. Contrastive learning is considered an algorithm that learns representations through the comparison of different input samples (i.e., anchor, positive, and negative samples). The advent of MoCo [41] and SimCLR [16] in **Computer Vision (CV)** has further promoted the development of contrastive learning, and MoCo and SimCLR have become the two most mainstream frameworks, widely adopted by subsequent models [11, 19, 30, 36, 118, 125, 134, 136]. The specific frameworks of MoCo and SimCLR are presented in Figure 2.

MoCo emphasizes that the number of negative samples is critical to improving representation learning and employs a momentum encoder to dynamically update negative examples. The momentum encoder relies on the query encoder to update its parameters; the momentum is set to a relatively large value (e.g., $m = 0.99$) because a larger momentum value results in better performance than a smaller value. In addition, MoCo involves the building of a dynamic dictionary for contrastive learning, in which the latest batch of samples is added to the queue and the oldest batch is removed from the queue in each iteration. The introduction of the queue decouples the dictionary size from the mini-batch size, allowing the dictionary to be larger than the mini-batch size to achieve much better performance. Specifically, a given augmented sample (also known as a query) generated via the DA strategy and a key in the queue are considered a positive pair if they are different views of the same instance and a negative pair if they are not.

SimCLR constructs positive pairs and combines multiple DAs to learn high-quality representations. In SimCLR, one data instance is first transformed by the DA module to form a positive pair, which is then fed into the encoder and the projector to extract semantic features. The projector is theoretically unnecessary because the encoder outputs can be directly used as metric representations. However, it has been shown that the projector in SimCLR [16] further compresses the redundant information contained in the encoder outputs, capturing more general representation information and improving contrastive learning efficiency. Therefore, the projector is usually added when using SimCLR as the contrastive learning framework. Compared with MoCo, SimCLR can be more easily implemented, but a large batch size of 8,192 is required to obtain relatively good performance.

Table 1. Mathematical Expression of Various Contrastive Loss Functions, in Which f Is the Hidden Representation of Sample, $m > 0$ Acts as a Radius of the Query, $s()$ Refers to the Metric Distance Function, and τ Is the Temperature Hyper-Parameter

Loss Type	Mathematical Expression
Pair loss [20]	$\mathcal{L}(x_i, x_j; f) = \mathbb{1}_{y_i=y_j}\|\|f_i - f_j\|\|_2^2 + \mathbb{1}_{y_i\neq y_j} max(0, m - \|\|f_i - f_j\|\|_2)^2$
Triplet loss [91]	$\mathcal{L}(x_i, x^+, x^-; f) = max(0, \|\|f - f^+\|\|_2^2 - \|\|f - f^-\|\|_2^2 + m)$
N-pair loss [96]	$\mathcal{L}(x_i, x^+, \{x_i\}_{i=1}^{N-1}; f) = -\log \frac{\exp(f^T f^+)}{\exp(f^T f^+) + \sum_{i=1}^{N-1} \exp(f^T f_i)}$
InfoNCE loss [80]	$\mathcal{L}(x_i, x^+, \{x_i\}_{i=1}^{N-1}; f) = -\log \frac{\exp(s(f, f^+)/\tau)}{\exp(s(f, f^+)/\tau)) + \sum_{i=1}^{N-1} \exp(s(f, f_i)/\tau)}$

In contrastive sentence representation learning, BERT and its variants typically serve as sentence encoders for feature learning. DisCo [117], however, involves the use of contrastive knowledge distillation [117] to obtain a more powerful student model from the larger teacher model sentence-T5 [75] as the sentence encoder. The sentence momentum encoder in MoCo is also a BERT-like PLM. The projector can be either a simple linear projection function or a nonlinear **Multi-Layer Perceptron (MLP)**. Most studies have adopted the MLP, as Chen et al. [16] demonstrated that the MLP could produce better results than simple linear projection function.

2.1.2 Contrastive Loss Function. The loss function, an indispensable component of contrastive learning, is the most significant difference between contrastive learning models and other neural network models. Representation learning models can be generally categorized into generative and discriminative models. The loss function of generative models is the metric distance between the inputs and generated data, and that of discriminative models is the distance between the labels and predicted results. The contrastive loss function is measured by the representation distance of the inputs in the projection space, where the projection distance can be the canonical L1-norm (Manhattan distance), the L2-norm (Euclidean distance), the dot product, cosine similarity, or bilinear similarity.

The design of the loss function is critical in contrastive learning, as it guides the training direction and objectives. As illustrated in Table 1, the contrastive loss function has evolved from the original pair loss [20], triplet loss [91], and N-pair loss [96] to the InfoNCE loss [80]. However, the objective of the contrastive loss function remains to maximize the distance between positive pairs and minimize the distance between negative pairs. **Mutual Information (MI)** loss is another frequently used loss function in representation learning, and MI loss based models learn representations by maximizing the MI between the inputs and their hidden vector representations [80, 101]. Oord et al. [80] showed that maximizing the lower bound on the MI loss is equivalent to minimizing the InfoNCE loss.

InfoNCE loss is the most widely used contrastive loss function and is regarded as a multi-label classification version of NCE (noise contrastive estimation) [39]. NCE is employed for discriminating data samples from noise samples and learning the distinctions between them to extract features from the data. Thus, InfoNCE loss uses categorical cross-entropy loss to classify the positive samples correctly from other negative samples. Assuming that we have a mini-batch of samples $\{x_1, x_2, \ldots, x_N\}$ and each sample x_i has one corresponding positive sample x_i^+, we can express the representations of these samples as $\{z_1, z_2, \ldots, z_N\}$ and $\{z_1^+, z_2^+, \ldots, z_N^+\}$. Let $S()$ denote the metric distance function that computes the representation distance between two inputs, and let τ be a temperature hyper-parameter. The InfoNCE loss for (x_i, x_i^+) with a mini-batch of N pairs is as

follows:

$$\mathcal{L}_i = -\log \frac{\exp(S(z_i, z_i^+)/\tau)}{\exp(S(z_i, z_i^+)/\tau) + \sum_{j=1, j\neq i}^{N} \exp(S(z_i, z_j)/\tau)}. \quad (1)$$

Both MoCo and SimCLR employ the InfoNCE loss as the contrastive loss function, but MoCo adopts the dot product as the distance function, whereas SimCLR uses the cosine similarity (L2-normalized dot product) as the distance function. The feature representations can be mapped into a unit hypersphere using cosine similarity, thereby improving training stability and model performance. The supervised contrastive loss function [52] incorporates the label information based on the loss function of SimCLR [16], and it allows for many positive samples per anchor sample (augmented samples and samples with the same class).

The temperature hyper-parameter τ also plays an important role in the InfoNCE loss. The empirical results from Wang and Liu [106] showed that InfoNCE loss allows for the self-discovery of hard negatives, and τ can regulate the degree of focus on hard negatives. The smaller the τ value, the greater the attention paid to the separation of anchor points from the hard negative samples. However, τ cannot be too small because the InfoNCE loss will degenerate to the triplet loss that only considers hard negatives as τ approaches 0. At present, almost all of the contrastive sentence representation models adopt $\tau = 0.05$ for model training. Moreover, a larger mini-batch size is usually beneficial for enhancing contrastive learning, as it partially increases the likelihood of hard negatives in negative samples.

2.2 Sentence Representation Models

Sentence representations, which encapsulate the semantic and syntactic features of a sentence in the form of dense vector embeddings, have been extensively studied. A straightforward solution is to use the composition of word embeddings, which considers that sentence representations can be composed by the average [71, 84] or weighted average [7, 89] of all of the word embeddings in the sentence. However, these methods fail to learn the information related to the word order and semantics of a whole sentence.

Deep neural network models such as **Recurrent Neural Networks (RNNs)** and **Convolutional Neural Networks (CNNs)** [23, 51, 55] allow for the automatic and sequential encoding of sentences into vector embeddings. However, RNN-based models lack transferability, and CNN-based models require careful setting of the convolution kernel size. Some studies [33, 42, 56, 65] have proposed constructing an encoder-decoder framework to learn sentence representations, in which the encoder maps sentences into fixed-length vectors, and the decoder predicts the surrounding sentences. An exception is the Quick-thought [65] framework, which regards the prediction of context sentences as a classification problem.

The emergence of Transformer [103] has facilitated the development of sentence representation learning. Transformer also adopts the encoder-decoder framework, but instead of using conventional CNNs and RNNs as the sentence encoders, the multi-head self-attention mechanism is employed to mine inter-word contextual information. For instance, the transformer-based conditional masked language model (CMLM) [127] learns representations through the recovery of masked tokens in a sentence conditioned on its contextual sentence representations. The sentence-level language model (SLM) [57] and the transformer-based sequential denoising auto-encoder (TSDAE) [108] capture robust sentence representations through the reconstruction of correct sentences from shuffled input sentences. Additionally, transformer-based PLMs such as BERT can be applied directly to learn sentence representations.

However, the vanilla BERT-like models suffer from poor sentence representations of the [CLS] token and underperform the average GloVe embeddings [84], attributable to the non-smooth and

anisotropic BERT token embedding distribution [58, 88]. The token embeddings of high-frequency words are clustered and close to the origin, whereas low-frequency words are sparsely distributed, resulting in a high similarity between any sentence pair. To ameliorate this issue, BERT-flow [58] and BERT-whitening [98] have been proposed to convert an anisotropic space into a smooth and isotropic space.

Several studies have also utilized labeled datasets to learn sentence representations. InferSent [6] leverages supervised **Stanford Natural Language Inference (SNLI)** datasets [10] to train a Siamese bidirectional long short-term memory (BiLSTM) network with a max-pooling layer output. Likewise, the universal sentence encoder (USE) [14] uses annotated data from the SNLI corpus to train the transformer network for representation learning. Reimers and Gurevych [88] proposed SBERT, which fine-tunes Siamese and triplet networks [91] on the combination of the SNLI [10] and **Multi-Genre Natural Language Inference (MNLI)** [114] datasets.

Motivated by the success of contrastive learning in CV, many models have employed contrastive learning to fine-tune BERT-like PLMs to learn sentence representations. More importantly, these contrastive learning models (see Section 3) outperform the aforementioned unsupervised models. Some of these models even show comparable performance to supervised sentence embedding models such as InferSent and USE and outperform the state-of-the-art SBERT.

Sentence representation learning has long been a research focus in NLP, possibly because sentences, as the basic unit of NLP tasks, can express more complete semantic information than individual words and can be more easily studied than the complex semantics of documents. Furthermore, sentence representations can be applied to almost all NLP tasks, including sentence classification [60, 61], sentiment analysis [74, 99, 135], machine translation [81], and semantic matching [37, 66].

3 CONTRASTIVE LEARNING METHODS FOR SENTENCE REPRESENTATIONS

More recently, there has been a surge in the development of approaches that employ contrastive learning to learn sentence representations in both self-supervised and supervised settings, all of which exhibit excellent performance. Moreover, unsupervised representation learning approaches, such as BYOL (Bootstrap Your Own Latent) [38], SimSiam [18], and Barlow Twins [131], all implicitly employ contrastive learning. Such studies are not included because they did not involve negative samples during model training and thus cannot be viewed as contrastive learning models.

This section first summarizes the contrastive learning frameworks used by various models and then presents a taxonomy based on the construction of positive pairs, negative samples, and external training data. The improved performance of these models is attributable to the combined effect of multiple factors, particularly the classification characteristics of the models. We provide a comprehensive overview of contrastive learning based sentence representation models and highlight their main features in Table 2.

3.1 Framework

The frameworks MoCo and SimCLR have been extensively applied to contrastive learning. The majority of contrastive sentence representation learning models use SimCLR as the contrastive learning framework rather than MoCo. For instance, contrastive learning based sentence representation models proposed in other works [22, 35, 36, 48, 49, 54, 70, 76, 92, 107, 117, 119, 125, 133, 134, 136, 138, 139, 142] employ SimCLR as their contrastive learning framework, whereas only a small number of models [11, 30, 59, 100, 118] adopt MoCo as their contrastive learning framework. The higher usage of SimCLR is attributable to its focus on obtaining more effective positive pairs, which favors the design of different models via the DA strategy [35, 36, 48, 54, 119, 125], whereas MoCo is less sensitive to the effects of positive pairs. In addition, the success of SimCSE [35], a

Table 2. Overview of Contrastive Learning Models for Sentence Representations

Model	Framework	Positive Samples	Negative Samples	Main Features
IS-BERT [137]	DIM	Global and local feature of the instance	Global and local features of other instances	MI loss function.
CERT [30]	MoCo	Back-translation	Samples in negative queue	Predict whether augmented sentences from the same sentence.
ESimCSE [118]		Dropout; Word repetition		Remove length feature between the positives and negatives.
PT-BERT [100]		Parameters difference between encoders		Map the positives and negatives into pseudo token embeddings of the same length and syntactic structure.
MoCoSE [11]		Dropout; FGSM		Explore how the negative queue length affects performance of contrastive learning.
AdCSE [59]		Parameters difference between encoders	Utilize adversarial training to produce hard negatives	High-quality negative representations help learn powerful sentence embeddings.
CLINE [105]	SimCLR	Synonym replacement	Antonym replacement	Only use one negative sample for representation learning.
DeCLUTR [36]		Two nearby segments in the same document	Text segments from other documents	Employ documents for sentence representation learning.
CLEAR [119]		Span deletion; Synonym replacement	In-batch negatives	Attempt various DA combinations.
ConSERT [125]		Token shuffling; Feature cutoff		Solve the collapse issue of BERT-derived sentence representations.
SG-OPT [54]		Parameters difference between encoders		Use the signal of fixed BERT to help tune another BERT.
Prompt-BERT [48]		Prompt-based augmentation		Avoid degradation caused by invalid BERT and token embedding bias.
SimCSE [35]		Dropout		Produces an excellent result with a simple augmentation method.
DCPCSE [49]		Dropout		Add multi-layer continuous prompts to the inputs, reducing training parameters.
USCAL [70]		Dropout; Adversarial examples		Use the gradient of contrastive loss to generate adversarial examples.
VaSCL [134]		Dropout	Use embeddings of anchors with optimal noise as hard negatives	Leverage the k-nearest in-batch neighbors of an instance to obtain the optimal noise.
MixCSE [138]		Dropout	Mix positive and negative features as hard negatives	Prove that hard negatives are essential in maintaining strong gradient signals.
CARDS [110]		Switch-case augmentation	Use text retrieval to find hard negatives	Combine the novel augmentation and hard negatives to yield a better result.
PairSupCon [133]		Entailment pairs of NLI datasets	Use negative samples with highest importance as hard negatives	Optimize a pairwise entailment and contradiction reasoning jointly.
DCLR [142]		Dropout	Extend in-batch negatives with noise-based negatives	Punish the false negatives to guarantee the uniformity of the representation space.
EASE [76]		Dropout; Entity	In-batch negatives and hard negative entities	Use Wikipedia entity for entity-aware contrastive learning.
DiffCSE [22]		Dropout	In-batch negatives	Generate the edited samples to help model discern the difference in sentences.
SNCSE [107]		Prompt-based augmentation		Introduce soft negative pairs to mitigate feature suppression.
MCSE [136]		Dropout; Sentence-image pairs		Use sentence-image pairs for multimodal contrastive learning.
ArcCSE [139]		Dropout; 20% masked anchor samples	In-batch negatives; 40% masked anchor samples	Add an angular m to the representations of positive pairs in InfoNCE loss.

contrastive sentence representation model built using the SimCLR framework, has inspired the development of numerous models [22, 76, 134, 136, 136, 138, 142] using SimCLR.

Info-Sentence BERT (IS-BERT) [137] uses the deep InfoMax (DIM) [44] framework and MI maximization for sentence representation learning, in which CNNs with different window sizes are used to produce local n-gram token embeddings that form positive pairs with the global sentence embeddings. Zhang et al. [139] discovered that the InfoNCE loss was insufficient to separate dissimilar sentences apart and pull sentences with similar semantics close. They then proposed the additive angular margin contrastive (ArcCon) loss, in which an additive angular margin of m is added between the hidden representations of positive pairs to make the model more tolerant and robust to noise. For a mini-batch of samples $\{x_1, x_2, \ldots, x_N\}$, with the assumption that h_i and h_i^* are two representations of sentence x_i and that cos () is the cosine similarity function, the ArcCon loss can be expressed as follows:

$$\mathcal{L}_{arc} = -\log \frac{\exp(cos(\theta_{i,i^*} + m)/\tau)}{\exp(cos(\theta_{i,i^*} + m)/\tau) + \sum_{j \neq i} \exp(cos(\theta_{i,j})/\tau)}, \tag{2}$$

in which angular $\theta_{i,j} = \arccos(\frac{h_i^T h_j}{||h_i|| \cdot ||h_j||})$.

3.2 Positive Pairs

Since one of the training goals of contrastive learning is to pull positive pairs close, the quality of positive pairs is crucial for noise-invariant sentence feature extraction. Nonetheless, contrastive learning heavily relies on DA to generate positive pairs. Therefore, the DA strategy greatly influences the effectiveness of representation learning. Text augmentation strategies for contrastive sentence representation learning can be broadly classified into token-level augmentation, sentence-level augmentation, document-level augmentation, PLM-based augmentation, and prompt-based augmentation. In the following, these text augmentation strategies are discussed in detail.

3.2.1 Token-Level Augmentation. This strategy involves the acquisition of augmented samples by making transformations at the token embedding layer of sentences or subword tokenization. For instance, ConSERT [125] leverages four text augmentation methods: token shuffling, token cutoff, feature cutoff [93], and dropout [43], and randomly selects two DAs to form positive pairs. All of these DA strategies work at the token embedding layer. ConSERT performs best on the STS task when positive pairs are generated by token shuffling and feature cutoff. The term *dropout* [43] refers to setting the elements in the token embedding matrix to zero with a certain probability. In particular, ConSERT shows that token frequency is the primary cause of poor BERT representations, whereas contrastive learning reshapes the original representation space of BERT. Contrastive learning with augmented and retrieved data for sentence embedding (CARDS) [110] develops a novel augmentation strategy, switch-case augmentation, to mitigate the token embedding bias of BERT-like PLMs. Substitution, division, fusion, and regrouping are four transformations of switch-case augmentation, which we illustrate with an example in Table 3. Switch-case augmentation is case sensitive and therefore cannot be applied to uncased BERT-like PLMs such as BERT-base-uncased and RoBERTa-base-uncased.

3.2.2 Sentence-Level Augmentation. All transformations that manipulate a sentence to produce another sentence while leaving the semantic meaning unchanged are considered in this review to be sentence-level text augmentation, which covers traditional text augmentation approaches such as synonym replacement and **Back-Translation (BT)**, and several random DA strategies. CLINE [105] generates positive and negative samples by substituting words with synonyms and

Table 3. Augmented Sentence Examples for Switch-Case Augmentation

Original	The book recommended is natural-istic. (Tokenization)
Case Switched	The book recommended is Natural-istic. (Substitution)
	The book recommended is Natural-is-tic. (Division)
	The book recommended is Naturalistic. (Fusion)
	The book recommended is Na-turalistic. (Regrouping)

antonyms from WordNet [72], respectively, and then minimizes the N-pair loss [96] of positive and negative samples and the original text to learn representations. CERT [30] employs BT to obtain augmented samples and then fine-tunes the BERT encoder by predicting whether the two augmented sentences are derived from the same sentence so that the encoder can capture global semantic features. Contrastive learning for sentence representation (CLEAR) [119] chose DA strategies such as word deletion, span deletion, span swap, and synonym replacement for contrastive learning. The experimental results demonstrated that different augmentations can obtain different sentence features and that mixed DAs are not always stronger than single DAs, whereas the combination of synonym replacement and span deletion achieves better overall performance.

3.2.3 Document-Level Augmentation. This text augmentation technique is typically applied to the learning of contrastive textual representations of long documents. For example, DeCLUTR [36] assumes that the two nearby textual segments in the same document are more likely to be semantically similar and can thus be regarded as a positive pair. It learns informative document representations by minimizing the distance between the embeddings of positive pairs.

3.2.4 PLM-Based Augmentation. Contrastive sentence representations are learned by fine-tuning BERT-like PLMs using contrastive learning. This type of text augmentation encodes the same sentence to produce positive pairs by leveraging the parameter differences or dropout mask between two BERT-like sentence encoders.

Parameters Difference Between Encoders. This augmentation strategy employs two encoders with different parameters to encode the same sentence, yielding two distinct representations as positive pairs. The self-guided contrastive learning for sentence representations (SG-OPT) [54] clones BERT into two copies: one with fixed parameters for computing the hidden representation of the intermediate layers, and the other for fine-tuning sentences to obtain the [CLS] representation. This way, two distinct hidden representations of the sentence are obtained: the intermediate layer representation and the [CLS] representation, which can be viewed as a positive pair. Contrastive tension [12] adopts a Siamese network-like architecture, in which two pretrained BERT encoders with the same structure but different parameters are used to encode the same sentence to create two sentence embeddings as the positive samples. Similarly, strategies involving the encoding of the same sentence, leveraging the parameters' difference between MoCo's gradient and momentum encoders to form a positive pair, have also been reported [59, 100]. In particular, PT-BERT [100] adds an extra pseudo-token embedding layer and an attention layer to the gradient encoder and momentum encoder to ensure that the generated positive and negative samples have the same length and syntactic structure, thereby eliminating the adverse effects of these differences.

Dropout Augmentation. Dropout augmentation can be seen as a minimal augmentation strategy that leverages the randomness of the dropout mask [97] in the fully connected layer and attention layers of transformer-based PLMs [27, 63]. Specifically, the same sentence is passed twice to the BERT-like sentence encoders to obtain two sentence embeddings as a positive pair. SimCSE [35] is the first contrastive sentence embedding model that applies dropout augmentation to

Table 4. Optional Templates for Sentence Representations

Relationship Prompts	[X] [MASK].
	[X] is [MASK].
	[X] mean [MASK].
	[X] means [MASK].
Prefix Prompts	This [X] means [MASK].
	This sentence of [X] means [MASK].
	This sentence of "[X]" means [MASK].
	This sentence: "[X]" means [MASK].

The last two templates in bold perform better and are used as prompt templates to represent sentences in the work of Jiang et al. [48] and Wang et al. [107].

create positive samples for sentence representation learning but yields a new state-of-the-art result. Because of the simplicity and superiority of the dropout augmentation strategy, it has been used to obtain positive instances to realize further improvements. The unsupervised contrastive adversarial learning (USCAL) model [70] first applies the dropout strategy to produce a positive pair and, then uses the gradient of dropout contrastive loss as adversarial noise to generate an adversarial example as another positive sample. Accordingly, the objective of USCAL is to minimize the contrastive loss between anchors and augmented examples (dropout) and that between anchors and adversarial examples. The motivation of ESimCSE [118] is that all positive pairs built by SimCSE have the same length, which could lead the model to misinterpret this as a distinguishing feature between positives and negatives. Therefore, a word repetition feature is applied to change the length of input sentences while keeping the semantics unchanged. The anchor sentence and its modified counterpart are then fed into the pretrained encoder with a random dropout to obtain positive views. DCPCSE [49] also leverages dropout augmentation to produce positive pairs, but it prepends multi-layer trainable dense vectors as continuous prefix prompts to the input sentences. During model training, DCPCSE only optimizes the deep continuous prefix prompts while freezing the parameters of pretrained BERT, reducing the complexity of parameter fine-tuning and tedious searching for handcrafted prompts. SupCL-Seq [92] adopts dropout augmentation to provide additional positive pairs for supervised sentence representation learning in addition to using samples with the same label as positive samples.

3.2.5 Prompt-Based Augmentation. This text augmentation technique is designed to mitigate the poor performance of the original BERT caused by the invalid BERT layer and token embedding bias. In particular, prompt-based augmentation [48, 107] aims to exploit the differences in prompt templates to obtain positive pairs. It first employs two discrete prompt templates to map the same sentence to the [MASK] token and then feeds this [MASK] token into BERT-like PLMs to produce two sentence embeddings as positive pairs. The commonly used discrete prompts can be found in Table 4, where [X] is used to place the input sentence, [MASK] represents the [MASK] token, and the hidden vector representation of the [MASK] token is viewed as the final sentence representation.

3.3 Negative Samples

Although positive pairs are important for the success of contrastive learning, negative pairs, particularly hard negatives, are equally significant [50, 77, 91, 106, 122, 141]. Much attention has thus been placed on the construction of hard negatives, but research on how to create effective negative samples to facilitate sentence representations is limited.

Table 5. Three Examples of Sentence Pairs with Different Relationships in the NLI Datasets

	ID	Sentence	Label
Premise	0	A person on a horse jumps over a broken down airplane.	
	1	A person is outdoors, on a horse.	*Entailment*
Hypothesis	2	A person is training his horse for a competition.	*Neutral*
	3	A person is at a diner, ordering an omelette.	*Contradiction*

The second hypothesis is neutral to the premise because "training" and "competition" are not reflected in the premise. The third hypothesis is a contradiction because it mentions a completely different event.

3.3.1 Hard Negatives. Hard negative samples are samples whose labels are different from the anchors but whose semantics are similar to the anchors such that it is difficult to distinguish them from the anchors. In the field of NLP, the methods frequently used to generate hard negatives mainly involve leveraging the label information of supervised datasets [35, 76, 117] and designing various algorithms [59, 110, 133, 134, 138].

Both supervised SimCSE [35] and Disco [117] employ the contradiction pairs of supervised **Natural Language Inference (NLI)** datasets as hard negatives, whereas the entity-aware contrastive learning of sentence embedding (EASE) model [76] leverages the Wikipedia entity supervision to obtain hard negatives. The NLI datasets are a combination of the MNLI dataset [114] and the SNLI dataset [10]. Each sample in the NLI datasets includes a sentence pair (premise-hypothesis) and a relationship label (entailment, neutral, or contradiction) for the pair, which can be illustrated using three examples in Table 5. PairSupCon [133] also adopts supervised NLI datasets as training corpus with entailment pairs as positive views. However, it leverages importance sampling to select negative samples with relatively high importance as hard negatives, which are constructed in an unsupervised manner. The negative samples in PairSupCon are other in-batch entailment pairs, as this model aims to jointly optimize pairwise entailment and contradiction reasoning to capture the high-level categorical semantic structure.

MixCSE [138], an unsupervised sentence representation learning approach, also highlights the significance of hard negatives. It extends SimCSE by mixing positive features and random negative features to produce hard negatives. Moreover, MixCSE shows that hard negatives are essential for keeping strong gradients and that randomly sampled negative examples are ineffective for contrastive sentence representations. AdCSE [59] was inspired by the success of AdCo [46] in CV. It utilizes adversarial training to train the negative sample queue in MoCo to produce hard negatives in an unsupervised manner. Because contrastive learning aims to minimize the contrastive loss by pulling the negatives apart from the positive samples, whereas the negative adversaries tend to confuse the neural network by maximizing contrastive loss, this adversarial training strategy leads to the generation of hard negatives.

VaSCL [134] utilizes a more generic technique for obtaining hard negatives. It first creates a neighborhood of an instance according to the k-nearest in-batch neighbors in the representation space. An instance-level contrastive loss is then defined to maximize the distance between instances and instances mixed with Gaussian noise while simultaneously minimizing the distance between instances mixed with Gaussian noise and neighborhood samples to obtain the optimal noise. The embeddings of the original instances with the optimal noise can be regarded as hard negatives of the original instances. Experimental results have indicated that VaSCL (which adopts dropout augmentation) outperforms SimCSE, verifying the effectiveness of the hard negatives. CARDS [139] retrieves the top k negative samples from the training corpus for each sample using text retrieval [120, 132], then computes the cosine similarity between the anchor and retrieved negative instance to screen out hard negatives. The introduction of such retrieved hard negatives greatly improves the model performance.

3.3.2 Effective Negative Samples. Beyond generating hard negatives, other operations can also be conducted on the negative samples to boost sentence representations. For instance, debiased contrastive learning of unsupervised sentence representations (DCLR) [142] assumes that the use of in-batch negatives may cause false negatives or negatives with anisotropy to be involved in representation learning, degrading the uniformity of the representation space. Therefore, virtual adversarial training [73] was first used to generate noise-based negatives to extend in-batch negatives, and the similarity scores between original sentences and their negative samples were used to design an instance weighting approach to punish false negatives. In the work of Cao et al. [11], a MoCo-style sentence embedding model, MoCoSE, that investigates how the negative queue length affects the performance of contrastive learning was proposed. Empirical results showed that there was an optimal range of historical information for the negative sample queue and that the negative samples near the middle of the queue performed better.

3.4 External Training Data

Different from positive pairs and negative samples generated on top of the training corpus, additional data are typically reconstructed based on specific tasks or are generated from carefully designed datasets. In addition, the inclusion of additional data can either help models capture subtle sentence features or provide effective training signals for learning more general semantic information.

Difference-based contrastive learning for sentence embeddings (DiffCSE) [22], a sentence embedding model inspired by equivariant contrastive learning [26] in CV, was designed to improve the model sensitivity to the differences in sentences. DiffCSE adds a new task to SimCSE [35], which employs ELECTRA [24] to produce edited sentences. The approach for generating edited sentences is similar to BERT's masked language modeling in that some words in the original sentences are first randomly masked, then words are generated at the location [MASK]. The discriminator in ELECTRA is used to learn the differences between the original input and the edited sentences, making the model sensitive to the difference between these sentences. DiffCSE can thus learn more robust representations by combining the contrastive loss on insensitive text augmentation (e.g., dropout) with the prediction loss on sensitive text transformation (e.g., edited sentences).

In SNCSE [107], researchers argued that most existing models suffer from feature suppression—that is, they fail to discriminate and decouple textual similarity from semantic similarity. To remedy this issue, the authors innovatively took the negations of original sentences as soft negative samples and measured the cosine similarity difference between positive pairs and soft negative pairs using the bidirectional margin loss. The texts of these soft negative samples were highly similar to the original text, but the semantics were completely different, which mitigates the problem of feature suppression. Additionally, SNCSE borrows the idea of PromptBERT [48], in which two discrete prompts are used to encode the same sentence to obtain a positive pair.

EASE [76] leverages Wikipedia entity supervision to learn sentence representations according to entity-aware contrastive learning loss between sentences and corresponding related entities, and the contrastive loss between sentences with dropout noise. Entity-aware contrastive learning treats sentences and their corresponding semantically related entities as positive pairs and collects a hard negative entity for each positive entity. Incorporating entity-aware contrastive learning provides rich training signals for sentence representations, as entities have been shown to be a powerful indicator of text semantics [32, 62, 123, 124].

In addition to Wikipedia entity supervision, multimodal contrastive learning of sentence embeddings (MCSE) [136] leverages both visual and textual information to learn sentence representations. In the work of Zhang et al. [136], the researchers combined SimCSE with a multimodal contrastive learning objective, in which a collection of sentence-image pairs were

Table 6. Details of the Dataset in the STS Tasks, Transfer Tasks, and Short Text Clustering Tasks

Dataset	Size	Task Type	Metrics	Source
		STS Tasks		
STS12	3.1k			
STS13	1.5k			
STS14	3.8k			
STS15	3.0k	Semantic similarity	Spearman corr.	Misc.
STS16	1.2k			
STS-B	8.6k			
SICK-R	9.9k			
		Transfer Tasks		
MR (**2**)	11k	Sentiment analysis		Movie reviews
CR (**2**)	4k	Sentiment analysis		Product reviews
SST-2 (**2**)	70k	Sentiment analysis		Movie reviews
SUBJ (**2**)	10k	Subjectivity/objectivity	Classification acc.	Movie review snippets
MPQA (**2**)	11k	Opinion polarity		Newswire
TREC (**6**)	6k	Question type		TREC
MRPC (**2**)	5.7k	Paraphrase detection		Web news
		Short Text Clustering Tasks		
Ag News (**4**)	8k	News title		News
Search Snippets (**8**)	12.3k	Search snippets		Web
Stack Overflow (**20**)	20k	Question title	Clustering acc.	Kaggle
Biomedical (**20**)	20k	Paper title		PubMed
Tweet (**89**)	2.5k	Social media content		Tweet
Google News (**152**)	11.1k	News title and snippets		Google

The bold numbers in the "Dataset" column denote the number of classes in the transfer tasks and the number of clusters in the short text clustering tasks, respectively. The size of each dataset refers to the approximate number of samples in the dataset.

used as training data to pull semantically related sentence-image pairs close and push unrelated pairs apart in the contrastive learning framework. Experimental results demonstrated that incorporating small-scale multimodal data into the text-only corpus enhanced the alignment of textual embedding and markedly improved the model performance.

4 EVALUATION TASKS FOR SENTENCE REPRESENTATIONS

In this section, we discuss typically used evaluation tasks for measuring the quality of sentence representation. The STS task is undoubtedly the most widely used method for assessing sentence representations and has been adopted by many sentence embedding models. As complementary tasks to STS tasks, some studies have used transfer tasks and short text clustering tasks to further illustrate the superiority of learned sentence representations. Specific datasets involved in these evaluation tasks are detailed in Table 6.

4.1 STS Tasks

STS tasks include the datasets STS 2012-2016 [1, 2, 3, 4, 5], STS-B (STS Benchmark) [13], and SICK-R (SICK-Relatedness) [68], which are the most frequently used tasks for evaluating sentence representations [12, 35, 36, 118, 119, 125, 133, 136, 137, 138, 142]. Datasets STS12-16 include only the test dataset, whereas STS-B and SICK-R include the training, test, and development sets. The development set of STS-B serves as the evaluation dataset for selecting the best checkpoint during model

training. Each sample in these datasets comprises a sentence pair and a human-annotated score ranging from 0.0 (different) to 5.0 (equivalent), where a higher score indicates a higher semantic similarity of the sentence pair.

To assess the quality of sentence presentation, we first calculate the semantic similarity of each sample pair in the test dataset and then compute the correlation coefficient between the calculated similarity and the human-annotated similarity. The default SentEval[3] usually applies a linear regressor on top of frozen sentence embeddings for STS-B and SICK-R and trains the regressor on the training sets of the two datasets. Most studies directly adopt the raw sentence embeddings for cosine similarity calculation. The correlation coefficient metric between the calculated cosine similarity (−1.0 to 1.0) and the gold similarity of the sentence pair (0.0 to 5.0) can be obtained using both Pearson's and Spearman's correlation. Reimers et al. [87] showed that Spearman's correlation, which measures ranking rather than actual scores, is more suitable for assessing sentence representations.

The aggregation method is also an important part of Spearman's correlation computation. The STS12-16 datasets have multiple subsets, and the results for these subsets can be collected via two approaches. The first approach involves concatenating all of the subsets and then calculating the overall Spearman's correlation, whereas the second approach involves separately computing the results for each subset and then averaging them. Most models [35, 59, 88, 100, 118, 125, 137, 138] adopt the first approach, as it combines data from different domains and brings the evaluation closer to the real-world environment.

4.2 Transfer Tasks

The transferability of sentence representations is the ability to capture informative semantic features and apply them to various tasks. It can be evaluated using the transfer tasks[4] [25], which comprise seven sentence classification tasks from diverse domains. These are the sentiment analysis tasks MR (Movie Review) [83], CR (Customer Review) [45], and SST-2 (Stanford Sentiment Treebank) [95]; the subjectivity/objectivity classification task SUBJ (Subjectivity) [82]; the opinion polarity classification task MPQA (Multi-Perspective Question Answering) [113]; the question-type classification task TREC (Text REtrieval Conference) [104]; and the paraphrase detection task MRPC (Microsoft Research Paraphrase Corpus) [28], which is a sentence-pair classification task that judges whether the sentence pairs capture semantic equivalence relationships.

For sentence classification tasks, a logistic regression classifier or an MLP with one hidden layer is trained using frozen sentence representations generated by various approaches. The classification accuracy is reported as a measure of the model's performance on a certain dataset, with higher classification accuracy indicating better performance.

4.3 Short Text Clustering Tasks

The ability of a model to encode high-level category information into the representations, which has only been considered in recent works [76, 133, 134], is also important in sentence representation evaluation. The short text clustering task is ideal for evaluating this capability, as it requires high-level semantic representation information. Short text usually contains only a few words. Thus, its vector representations tend to be quite sparse, so clustering semantically similar texts together could be difficult. Consequently, only the model that learns high-level categorical structures can cluster similar sentences in the representation space.

[3] https://github.com/facebookresearch/SentEval.

[4] Transfer tasks are different from transfer learning. They are actually standard evaluation tasks for the quality of sentence representations. We follow this expression of "transfer tasks" to be consistent with statements in works such as SimCSE, MixCSE, EsimCSE, and DiffCSE.

Search Snippets [85], Stack Overflow [121], Biomedical [121], AgNews [86], Tweet [129], and Google News [129] are some of the most extensively used short text clustering datasets. These datasets are collected from Web search engines, news articles, question titles (Kaggle[5]), biomedical databases (PubMed), and social media (Tweet). The average sentence length of these datasets ranges from 6 to 28, and the number of clusters in these datasets has been determined in some works [85, 86, 121, 129] using several advanced algorithms. We cluster the sentence embeddings of these short text datasets[6] using the well-known algorithm k-means [64, 67] owing to its simplicity. The clustering accuracy is reported to measure the quality of learned sentence representations on these datasets.

4.4 Connection to Downstream Applications

The preceding evaluation tasks are also downstream applications of sentence representations. The STS tasks, which calculate the semantic similarities between sentence pairs, can be expanded to applications such as semantic matching [37, 66]. The transfer tasks are sentence classification tasks from diverse domains. The clustering of short texts into groups of similar texts also plays a critical role in numerous real-world applications such as topic discovery [53], trend detection [69], and recommendation [9]. Additionally, sentence representations can be used for machine translation, question answering, and almost all NLP tasks. Sentence representation is the basic processing unit of NLP tasks, and thus its quality extensively influences a variety of downstream tasks.

5 EXPERIMENT

Experiments were conducted on the evaluation tasks discussed in Section 4. Without loss of generality, we selected some representative models for the experiments. We ran experiments with the models on the STS, transfer, and short text clustering tasks, and comprehensively analyzed the quantitative results. Finally, we compared the alignment and uniformity values of these models on the STS-B development set, and the t-SNE visualization results of Stack Overflow representation obtained using these models.

5.1 Training Setup

5.1.1 Training Models. We used SimCSE as the baseline model and selected several representative models according to their most salient features for experiments. ArcCSE [139] and CARDS [139] were not included because the code for ArcCSE is not publicly available, and CARDS cannot be applied to uncased BERT-like PLMs owing to its switch-case augmentation property. The experimental models were selected according to the generation of positive pairs, negative samples, and additional training data:

- *Models that focus on the generation of positive pairs*: We first constructed the new models "SR-BERT," "BT-BERT," and "TSFC-BERT" based on the SimCLR framework; the models were named according to the employed DA techniques, namely synonym replacement (SR), BT, and token shuffling and feature cutoff (TSFC). We changed the framework of ESimCSE [118] from MoCo to SimCLR and renamed the model "EsimCSE-SimCLR." Moreover, the models SimCSE [35], DCPCSE [49], PT-BERT [100], and PromptBERT [48] were also considered. The dropout was closed during the application of DA TSFC to pretrained BERT-like models.

[5]Question title from Kaggle: https://www.kaggle.com/competitions/predict-closed-questions-on-stack-overflow/data?select=train.zip.

[6]Experimental short text clustering datasets can be found at https://github.com/rashadulrakib/short-text-clustering-enhancement/tree/master/data.

- *Models that focus on the construction of negative samples*: AdCSE [59] and MixCSE [138] were chosen for experiments to indicate the benefits of hard negatives, whereas the model DCLR [142] was selected to illustrate the significance of noise-based negative adversarial samples.
- *Models that focus on exploring additional data*: DiffCSE [22], SNCSE-Dropout, EASE [76], and MCSE [136] were chosen to test whether the addition of extra training data to SimCSE can help the model learn more general semantic features. SNCSE-Dropout refers to a variant of the SNCSE [107] model that uses dropout augmentation rather than prompt-based augmentation. This model can more effectively illustrate the effect of soft negative samples on contrastive sentence representations and aligns with the DA strategies of the other three models.
- *A model that combines positive pairs, hard negatives, and additional training data*: As stated previously, positive pairs, negative samples, and external training data considerably influence the quality of contrastive sentence representations. Thus, we constructed a new model, "MixCSE-DiffCSE," by integrating MixCSE and DiffCSE.
- *Models that do not involve contrastive learning*: These models include GloVe embedding[7] [84] and the vanilla BERT model [27]. The post-processing methods BERT-flow [58] and BERT-whitening [98] were also considered for unsupervised comparison.

5.1.2 Training Details. All contrastive learning based sentence representation models were trained on 1 million randomly sampled unlabeled sentences from the English Wikipedia.[8] BERT-flow and BERT-whitening were trained on the unsupervised NLI datasets, whereas GloVe embeddings and vanilla BERT were pretrained on a large-scale corpus. Our entire empirical implementation was based on the Hugging Face Transformers library[9] [115]. The training was performed using the uncased pretrained BERT_{base} model (110 million parameters) as the sentence encoder. The model was fine-tuned with the contrastive learning objective. We reproduced the evaluation results using the code provided in the original papers, except the codes for the models SR-BERT, BT-BERT, TSFC-BERT, ESimCSE-SimCLR, and SNCSE-Dropout were rewritten on the basis of SimCSE, and MixCSE-DiffCSE was rewritten based on MixCSE and DiffCSE. All experiments were implemented on an NVIDIA 3060 with the PyTorch version of 1.11.0 + CUDA 11.3.

SR-BERT, BT-BERT, TSFC-BERT, ESimCSE-SimCLR, SNCSE-Dropout, and MixCSE-DiffCSE adopted the same batch size and learning rate as the unsupervised SimCSE. For the remaining models, the training parameters such as batch size and learning rate were consistent with the corresponding original papers, indicating that these models were trained with different parameter settings. We trained all contrastive learning models for one epoch, except for DiffCSE (which was trained for two epochs), as well as DCLR and MCSE (which were trained for three epochs). We evaluated these contrastive learning models every 125 training steps (250 training steps for SimCSE and EASE) on the development set of STS-B and obtained the best checkpoints for the final evaluation on the test set. This appears to violate the training parameter consistency criterion. The optimal results of these models were obtained according to the relevant parameters. Table 7 shows the specific batch size and learning rate adopted by the models during training. The temperature parameter τ was set to 0.05 in all experiments involving contrastive learning.

[7] https://huggingface.co/sentence-transformers/average_word_embeddings_glove.840B.300d.
[8] https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse/resolve/main/wiki1m_for_simcse.txt.
[9] https://github.com/huggingface/transformers.

Table 7. Batch Size and Learning Rate of Experimental Models

Model	Batch Size	Learning Rate
BERT-flow, BERT-whitening	16	2e-05
SR-BERT, BT-BERT, TSFC-BERT, UnsupSimCSE, ESimCSE-SimCLR, AdCSE, MixCSE, SNCSE-Dropout, MCSE, MixCSE-DiffCSE	64	3e-05
EASE	8	3e-05
DiffCSE	64	7e-06
DCLR	128	3e-05
PromptBERT	256	1e-05
DCPCSE	256	3e-02

5.2 Quantitative Results and Analysis

The experiments were conducted on STS tasks, transfer tasks, and short text clustering tasks, the details of which are given in Section 4. The primary purpose of sentence representations, according to Gao et al. [35] and Reimers and Gurevych [88], is to group semantically similar sentences. Hence, the results on the STS tasks were used for the main comparison, with the results on the transfer tasks and short text clustering tasks serving as supplements. Tables 8, 9, and 10 present the reproduced results of various sentence embedding models. The reproduced results may differ from those reported in the original papers owing to differences in GPU and CUDA versions during training. However, because all reproduced results were obtained in the same experimental environment, the comparison of experimental results was relatively fair.

The manner in which the BERT sentence representations are represented is also important, and there are primarily five pooling methods. The first is the [CLS] representation, in which the representation of the [CLS] token is used as the final sentence representation [49, 138]. The second is the [CLS] representation without MLP, in which an MLP layer was kept over [CLS] during training but removed during testing. This representation method was first adopted in the work of Gao et al. [35] and produced better results than the [CLS] representation in unsupervised scenarios. Consequently, this representation method has also been adopted in several models [22, 59, 100, 107, 138, 139, 142]. The third frequently used method is the "avg." representation [48, 76], which involves taking the average embeddings of the last layer of BERT. The fourth is the "first-last-avg." representation, which is the average embeddings from the first and last layers of BERT [88]. The fifth is the "last-2-avg." representation, which is the average embeddings from the last two layers of BERT [58, 98, 125].

5.2.1 Performance Evaluation on the STS Tasks. Table 8 shows the evaluation results on seven STS datasets, with Spearman's correlation as the evaluation metric. The evaluation results provide several findings.

Contrastive learning boosted the quality of learned sentence representations. Compared with previous methods such as average GloVe embedding and post-processing methods BERT-flow and BERT-whitening, almost all contrastive sentence embedding models exhibited substantial performance improvements, confirming the effectiveness of contrastive learning.

Prompt-based augmentation outperformed other text augmentation strategies. In the framework of contrastive learning, sentence representation models with PLM-based augmentation outperformed models with token-level and sentence-level augmentation. The dropout augmentation adopted by SimCSE [35] features the simplest mechanism; however, it yielded a moderately good result. In

Table 8. Sentence Representation Performance on the STS Tasks (Spearman's Correlation ×100%)

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Self-Supervised Models								
GloVe embeddings (avg.)	57.48	70.99	60.70	70.85	63.84	60.91	54.81	62.80
BERT_{base} ([CLS])	21.54	32.11	21.28	37.89	44.24	20.29	42.42	31.40
BERT_{base} (first-last-avg.)	39.69	59.37	49.67	66.03	66.19	53.88	62.06	56.70
BERT_{base} (avg.)	30.87	59.89	47.73	60.29	63.73	47.29	58.22	52.57
BERT-flow$^{\heartsuit}$	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT-whitening	61.68	65.72	66.05	75.16	73.21	68.29	63.65	67.68
SR-BERT$^{\diamondsuit}$	59.98	69.45	61.45	73.49	70.04	67.92	68.96	67.33
BT-BERT$^{\diamondsuit}$	69.52	77.98	69.26	79.76	75.38	73.10	70.67	73.67
TSFC-BERT$^{\diamondsuit}$	66.92	80.13	71.77	81.51	77.48	77.64	66.89	74.62
SimCSE	66.84	78.93	72.47	80.50	78.27	76.94	71.72	75.10
ESimCSE-SimCLR$^{\diamondsuit}$	68.95	82.05	75.28	81.82	78.44	78.38	69.30	76.32
DCPCSE	71.30	82.75	73.77	81.93	78.58	77.81	69.55	76.53
PT-BERT	68.72	82.46	73.96	81.42	77.99	77.73	70.90	76.17
PromptBERT	69.67	83.33	75.13	**83.86**	78.78	**80.86**	69.78	77.34
AdCSE	67.51	81.64	72.76	80.87	78.78	76.48	72.48	75.79
MixCSE	71.03	82.06	75.02	82.54	79.82	79.47	70.48	77.20
DCLR	66.95	80.45	73.37	81.58	77.32	77.86	71.26	75.54
DiffCSE	69.14	82.87	74.27	82.85	80.10	78.93	71.08	77.03
SNCSE-Dropout$^{\diamondsuit}$	67.70	81.96	74.30	81.95	78.55	77.87	**73.66**	76.57
SNCSE	69.31	**84.21**	**76.87**	82.35	**80.51**	80.45	73.61	**78.19**
EASE$^{\clubsuit}$	70.01	81.81	72.72	82.21	78.47	79.02	68.35	76.08
MCSE$^{\spadesuit}$	70.78	81.10	74.30	83.58	78.30	79.76	71.52	77.05
MixCSE-DiffCSE†	**71.75**	82.77	76.19	83.48	78.61	79.63	70.73	77.59

$\heartsuit$: Results from Gao et al. [35]. $\diamondsuit$: Results were obtained by retraining our rewritten SimCSE-based code. $\clubsuit$: Results were obtained by reproducing the checkpoint provided in the original paper. $\spadesuit$: MCSE was trained with five random seeds, and the means were reported. †: The combination of MixCSE and DiffCSE is a new model that considers the positive pairs, hard negatives, and external training data.

contrast, the prompt-based augmentation proposed by PromptBERT achieved the best result among the augmentation strategies, with an average STS performance of 77.34%. One hypothesis is that during the fine-tuning of large PLMs for contrastive learning, using well-designed DA techniques usually provides more benefits in capturing semantic information.

Enhancing negative samples improved contrastive learning. The introduction of hard negatives boosted the contrastive learning performance, compared with the SimCSE results. In particular, MixCSE yielded an average STS of 77.2%, which was 2.1% higher than that of SimCSE, indicating the effectiveness of hard negatives generated through the combination of positive and random negative features. Although the performance improvement of DCLR over SimCSE was negligible, it demonstrates the utility of noise-based negative samples.

MCSE outperformed the other three models that incorporate external data. DiffCSE, SNCSE-Dropout, EASE, and MCSE are contrastive learning based sentence representation models that add edited sentences, soft negative samples, Wikipedia entity supervision, and multimodal sentence-image pairs to SimCSE for model training. The experimental results showed that MCSE, which incorporates multimodal sentence-image pairs for multimodal contrastive learning, outperformed the other three models.

Table 9. Sentence Representation Performance on the Transfer Tasks
(Classification Accuracy ×100%)

Model	MR	CR	SUBJ	MPQA	SST-2	TREC	MRPC
GloVe embeddings (avg.)	75.82	78.38	89.18	84.90	78.91	82.40	71.42
$BERT_{base}$ ([CLS])	79.66	83.68	93.96	84.70	85.17	83.00	68.81
$BERT_{base}$ (first-last-avg.)	81.04	**86.84**	94.99	89.45	85.45	89.80	74.03
$BERT_{base}$ (avg.)	81.52	86.73	95.22	87.75	85.94	**90.60**	73.68
SR-BERT$^{\diamond}$	79.90	84.96	94.57	88.66	84.57	89.60	74.55
BT-BERT$^{\diamond}$	78.40	85.34	94.43	89.11	83.60	84.15	74.43
TSFC-BERT$^{\diamond}$	77.52	81.80	93.22	88.84	81.82	86.20	74.78
SimCSE	81.40	85.80	94.31	88.36	84.84	88.40	74.49
ESimCSE-SimCLR$^{\diamond}$	**81.79**	86.57	94.62	88.76	85.50	85.60	73.57
DCPCSE	77.80	83.89	92.15	87.98	81.71	76.20	73.51
PT-BERT	81.27	85.93	94.59	89.52	86.33	88.00	74.61
PromptBERT	80.42	86.18	93.51	89.05	84.35	87.60	75.19
AdCSE	80.76	85.56	94.68	89.10	85.17	87.00	73.74
MixCSE	81.12	85.09	94.40	88.73	85.56	85.80	74.61
DCLR	81.12	86.69	94.98	**89.72**	**86.35**	84.15	74.95
DiffCSE	81.65	86.10	95.01	89.17	86.05	85.60	**76.41**
SNCSE-Dropout$^{\diamond}$	80.09	85.41	94.53	89.10	85.06	87.20	73.51
EASE$^{\clubsuit}$	70.18	84.29	**99.60**	88.38	85.39	84.40	75.42
MCSE$^{\spadesuit}$	81.22	86.61	94.55	89.06	85.92	89.08	74.82

$\diamond$: Results were obtained by retraining our rewritten SimCSE-based code. $\clubsuit$: Results were obtained by reproducing the checkpoint provided in the original paper. $\spadesuit$: MCSE was trained with five random seeds, and the means were reported.

SNCSE outperformed the other models. With an average STS of 78.19%, SNCSE achieved the best results among all of the models, as well as the best results on the STS13, STS14, and STS16 datasets. The use of prompt-based sentence embeddings and soft negative samples possibly contributed to the outstanding performance of SNCSE. The effectiveness of prompt-based augmentation is demonstrated by the excellent performance of PromptBERT. The effect of soft negative samples can be illustrated by the performance of SNCSE-Dropout, which outperformed SimCSE by about 1.4%.

Vanilla BERT underperformed average GloVe embeddings in sentence representations. The three vanilla BERT models based on different pooling methods underperformed the average GloVe embeddings (non-contextualized embeddings trained with a simple model). The experimental results confirm the finding in the work of Li et al. [58] and Reimers and Gurevych [88]. In addition, both BERT-flow and BERT-whitening significantly improved the performance of vanilla BERT.

Combining three salient factors that influenced the quality of contrastive sentence representations could lead to better results. As MixCSE combines dropout augmentation and hard negatives, and DiffCSE combines dropout augmentation and external training data, the new model MixCSE-DiffCSE is a mixture of positive pairs, hard negatives, and additional training data. According to the experimental results at the bottom of Table 8, the average STS performance of MixCSE-DiffCSE was 77.59%, which was better than those of MixCSE (77.20%) and DiffCSE (77.03%), indicating that combining these three salient factors is likely to produce even better results.

5.2.2 Performance Evaluation on the Complementary Tasks. Tables 9 and 10 present the evaluation results on the seven sentence classification datasets from different domains and on the

Table 10. Sentence Representation Performance on the Short Text Clustering Tasks (Clustering Accuracy ×100%)

Model	Ag News	Search Snippets	Stack Overflow	Bio-medical	Tweet	Google News	Avg.
GloVe embeddings (avg.)	79.71	72.51	37.10	33.02	49.40	62.22	55.66
BERT_{base} ([CLS])	27.62	17.30	8.38	10.85	14.20	13.77	15.35
BERT_{base} (first-last-avg.)	**86.28**	69.87	33.90	34.36	49.33	67.18	56.82
BERT_{base} (avg.)	79.59	64.19	21.59	32.54	44.57	61.82	50.72
SR-BERT$^{\diamond}$	84.60	**74.44**	23.53	31.25	51.12	65.78	55.12
BT-BERT$^{\diamond}$	78.89	70.09	39.26	29.64	52.56	65.13	55.93
TSFC-BERT$^{\diamond}$	75.37	65.52	**71.18**	**40.14**	56.04	68.98	62.87
SimCSE	76.41	67.37	59.38	33.61	53.55	66.13	59.41
ESimCSE-SimCLR$^{\diamond}$	80.23	66.48	70.89	36.04	53.81	66.50	62.33
DCPCSE	76.30	55.60	61.61	36.93	50.44	65.13	57.67
PT-BERT	76.80	57.94	63.32	35.50	51.83	66.81	58.70
PromptBERT$^{\heartsuit}$	–	–	–	–	–	–	–
AdCSE	72.80	57.60	70.81	35.69	51.98	65.97	59.14
MixCSE	78.27	73.31	66.49	36.57	54.52	67.75	62.82
DCLR	78.78	55.63	59.64	35.15	51.80	66.28	57.88
DiffCSE	81.97	73.34	63.86	37.70	53.37	67.10	62.89
SNCSE-Dropout$^{\diamond}$	74.04	44.50	70.64	36.78	51.72	67.11	57.47
EASE$^{\clubsuit}$	85.40	72.31	68.16	35.94	**56.78**	**69.40**	**64.67**
MCSE$^{\spadesuit}$	79.98	64.43	60.58	36.22	53.64	66.58	60.24

$\diamond$: Results were obtained by retraining our rewritten SimCSE-based code. $\clubsuit$: Results were obtained by reproducing the checkpoint provided in the original paper. $\spadesuit$: MCSE was trained with five random seeds, and the means were reported. $\heartsuit$: The dimension of PromptBERT representations does not match the dimension of corresponding labels, we cannot perform k-means clustering and thus short text clustering.

six short text clustering datasets. The evaluation results were analyzed, and several trends were observed.

Overall performance on transfer tasks outperformed that on short text clustering tasks. The average classification accuracy of these sentence embedding models remained at about 85% in the seven transfer tasks. In six short text clustering tasks, the average clustering accuracy of these sentence embedding models remained at around 60%. Extracting semantic features from short text is rather difficult, and the short text clustering tasks contained more clusters (e.g., the Google News dataset had 152 clusters), whereas sentence classification tasks were a binary classification problem, which explains the preceding results.

Contrastive learning improved performance on most transfer tasks. The performance of BERT and contrastive learning based models on seven transfer tasks was compared. The vanilla BERT model based on "first-last-avg." and "avg." representation achieved higher performance than contrastive learning baselines on the CR and TREC datasets. The relatively small size of the CR and TREC datasets affected the training of linear classifiers based on frozen sentence representations and may have influenced the evaluation results of the contrastive learning baselines. However, the contrastive learning baselines achieved higher performance on the MR, SUBJ, MPQA, SST-2, and MRPC datasets. Thus, contrastive learning is useful for most transfer tasks.

Contrastive learning improved performance on most short text clustering tasks. The performance of BERT and contrastive learning-based models on six short text clustering tasks was compared. We observed that contrastive learning models perform better than vanilla BERT on Search Snippets,

Stack Overflow, Biomedical, Tweet, and Google News datasets, whereas the "first-last-avg." representation of BERT performs better on the Ag News dataset. Moreover, the average performance of BERT's [CLS] representation on the six short text clustering tasks was 15.35%, significantly lower than the GloVe average embeddings of 55.66%.

Model performance on the short text clustering datasets was unevenly distributed. The experimental models exhibited moderately good results on the Ag News and Google News datasets, with overall clustering accuracies of around 80% and 66%, respectively. However, they performed poorly on the Biomedical and Tweet datasets, with overall clustering accuracies of about 35% and 50%, respectively. Additionally, the performances of these models varied considerably on the Search Snippets and Stack Overflow datasets. In particular, we noticed that almost all of the models exhibited the worst performance on the Biomedical dataset, possibly because the Biomedical dataset differed significantly from the Wikipedia training corpus.

5.3 Alignment and Uniformity

To more effectively compare the performances of the aforementioned models, we utilized a contrastive representation analysis tool from Wang and Isola [109], in which alignment and uniformity were employed as general metrics to measure the quality of learned representations. Alignment measures the expected distance between the embeddings of positive pairs. Uniformity measures how well the embeddings of instances are uniformly distributed. Lower values of both uniformity and alignment represent better model performance. Following the setting in SimCSE [35], we took the sentence pairs in the STS-B development set with a similarity score higher than 4.0 as P_{pos} and all of the development datasets of STS-B as P_{data}. We denote as $f(x)$ the normalized representation of instance x. Then, alignment and uniformity can be defined as follows:

$$\mathcal{L}_{align} = \mathbb{E}_{(x,x^+)\sim P_{pos}} ||f(x) - f(x^+)||^2, \tag{3}$$

$$\mathcal{L}_{uniform} = \log \mathbb{E}_{(x,y)\sim P_{data}} e^{-2||f(x)-f(y)||^2}. \tag{4}$$

Figure 3 depicts the alignment and uniformity of various sentence embedding models and their average results across the seven STS tasks. We observed the following results:

- Pretrained BERT embedding models exhibited competitive alignment but poor uniformity (i.e., the embedding space is anisotropic).
- BERT-flow and BERT-whitening substantially enhanced the uniformity of vanilla BERT and demonstrated the effectiveness of the flow-based model and whitening operation; however, both models performed poorly in terms of alignment.
- Most contrastive sentence embedding models significantly improved the uniformity of pretrained BERT embeddings while maintaining good alignment.
- Among all models, PromptBERT and DiffCSE exhibited the best alignment and the worst uniformity among all models, possibly because PromptBERT adopted the hidden vector representation of [MASK] to eliminate the effects of invalid BERT layers and token embedding bias, whereas DiffCSE learned the difference between sentences through using edited sentences.
- DCPCSE greatly enhanced the uniformity but performed poorly in terms of alignment. This may be caused by the absence of pretrained BERT during the model training.
- MixCSE achieved the best overall alignment and uniformity, attributable to the combined effects of dropout augmentation and hard negatives.

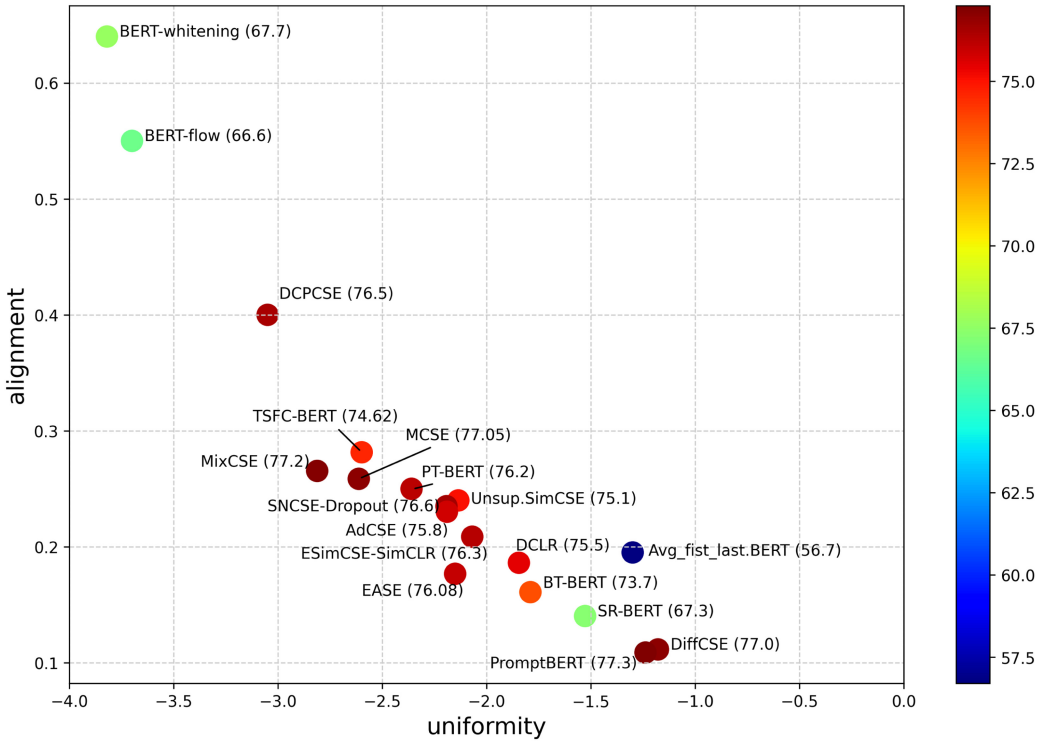

Fig. 3. Alignment and uniformity for different sentence representation methods measured on the STS-B development set. The color of points represents average performance on the STS tasks.

- SNCSE-Dropout exhibited nearly the same alignment and uniformity as unsupervised SimCSE, possibly because they shared the same positive and negative samples and the same model framework.
- The performance of these models on the STS tasks was generally consistent with the overall performance with respect to alignment and uniformity.

5.4 Visualization of Sentence Representations

Visualizing sentence embeddings is useful for understanding how well these models learn from sentences. To be considered a high-quality representation, embeddings for sentences with the same semantics should be close to each other, and meanwhile, embeddings for sentences with different semantics should be pulled apart. To visualize high-dimensional vector embeddings, a widely used dimension reduction technique, t-SNE [102], is employed to transform high-dimensional vectors into a two-dimensional vector space.

Following the experimental setting in PairSupCon [133], the experimental dataset for our t-SNE visualization was the short text clustering dataset Stack Overflow, which is a subset of Kaggle, covering 20,000 question titles with 20 clusters. We chose Stack Overflow as the experimental dataset because the model performance on the Stack Overflow dataset can better reflect the models' overall average performance on the six short text clustering tasks.

The visualization of Stack Overflow representations using t-SNE is shown in Figure 4. The [CLS] representation from BERT exhibited the worst performance and failed to achieve category clustering, with the "avg." representation of BERT performing slightly better than the [CLS]

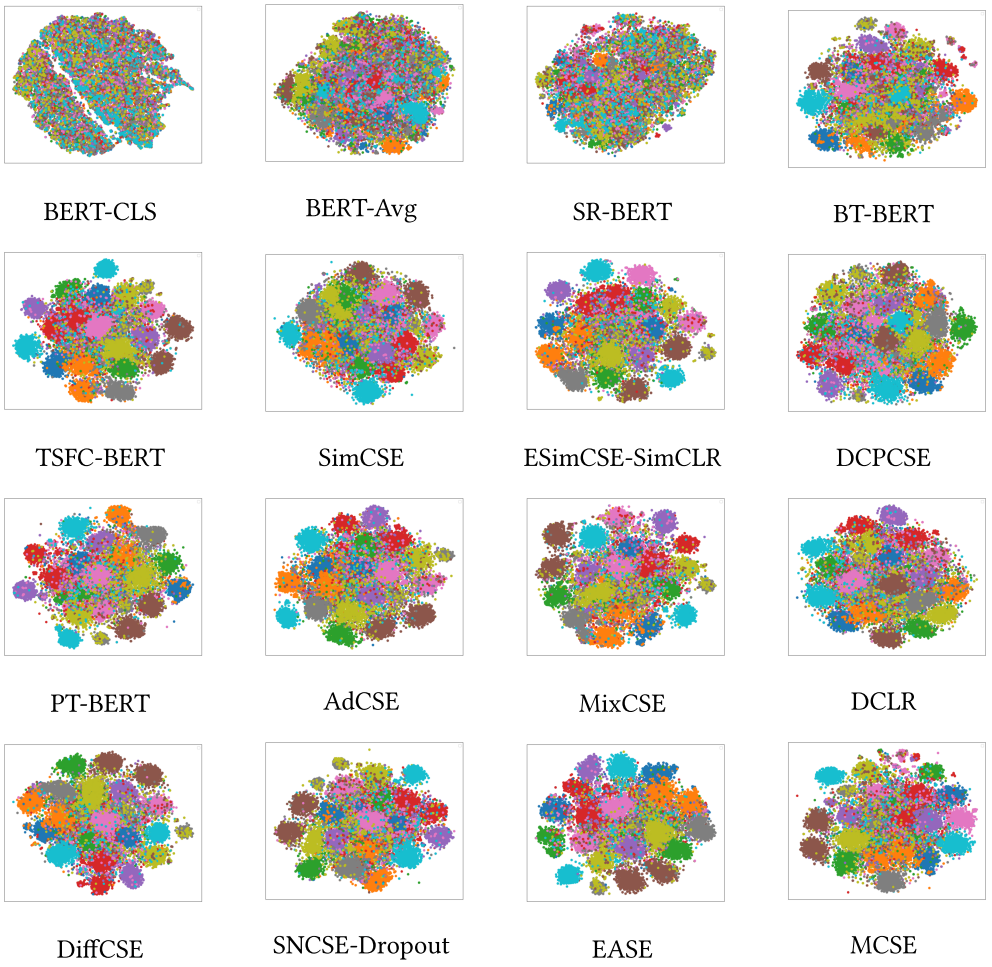

Fig. 4. The t-SNE visualization of the Stack Overflow representation using a series of models.

representation. SR-BERT and BT-BERT using simple DA techniques to generate positive pairs outperformed vanilla BERT but underperformed models using well-designed DA techniques. Additionally, all of the contrastive learning based sentence representation models were superior to vanilla BERT. In particular, TSFC-BERT demonstrated a strong clustering quality. Improved models based on SimCSE, such as ESimCSE-SimCLR, MixCSE, and EASE, showed a more powerful clustering capacity than SimCSE. Overall, the experimental results agree with our expectations that contrastive learning is beneficial to BERT-derived sentence representation learning and that constructing high-quality positive pairs, negative samples, and additional training data allows the model to learn high-level category structure information.

6 FUTURE RESEARCH DIRECTIONS

The introduction of contrastive learning has greatly contributed to the development of sentence representation learning, and there is still room for further exploitation and exploration. In this section, we discuss the areas that can be further investigated.

6.1 Data Augmentation

6.1.1 Designing a DA Strategy. According to recent studies on representation learning [16, 35, 36, 41, 139], contrastive learning is expected to be one of the standard paradigms for sentence representation learning. Positive pairs, as the crucial part of contrastive learning, have resulted in the development of a line of text augmentation strategies. These augmentation strategies comprise traditional text augmentations, such as BT, synonym replacement, random deletion, and swap, as well as well-designed text augmentations such as dropout augmentation and prompt-based augmentation. In particular, prompt-based augmentation achieves significant performance improvements by combining prompts to represent sentence embeddings. Thus, further research could follow this line to develop novel text DA strategies.

In addition, we observed that various DAs learned different sentence features. For instance, SR produced the best results in TREC and Search Snippets, dropout with word repetition augmentation yielded the highest classification accuracy in MR, and TSFC yielded the highest clustering accuracy in Stack Overflow and Biomedical. In other words, some specific augmentation strategies were particularly effective for certain downstream tasks, as discussed in the work of Wu et al. [119]. Therefore, designing a task-specific DA will be helpful for extracting specific sentence features.

6.1.2 Theoretical Explanation Lag Behind Empirical Novelties. Although DA strategies are beneficial for contrastive learning, and the effectiveness of DAs for contrastive sentence representation learning has been proved by a large number of experiments [30, 35, 105, 119, 125], few studies have theoretically explained the mechanism of DA in contrastive learning. Recent works [47, 112] have investigated and analyzed the feature learning process of contrastive learning in CV, and how DA helps boost the performance of contrastive learning. These studies, however, have focused on images, and the image data samples were represented by a sparse coding model [78, 79] or a spiked covariance model [8, 128]. Similar theoretical research in NLP is nonexistent owing to the discrete nature of texts (i.e., the text is a discrete variable rather than a continuous variable, so it is difficult to design an appropriate function to express text). Researchers could commence such theoretical studies by considering how the augmented samples generated by a specific text DA affect the performance of contrastive learning.

6.2 Negative Sampling

6.2.1 Hard Negative Mining. The importance of hard negatives to the improvement of contrastive learning has been emphasized by many works [50, 106, 122]. In the field of NLP, a straightforward solution is to use the contradiction pairs of the NLI datasets as hard negatives. Unfortunately, no work has explored the use of other supervised text datasets to obtain hard negatives.

Algorithms that are carefully designed to generate hard negatives are characterized by low generality or transferability, and most of them require cumbersome computation. As discussed in Section 3, PairSupCon [133] requires importance sampling, VaSCL [134] needs to obtain the optimal noise from the top-k neighborhood, AdCSE [59] requires external adversarial training, and CARDS [139] needs to leverage text retrieval to choose the top-k negatives before selecting the hard negatives using the cosine similarity between sentence pairs. These limitations restrict the application of hard negatives to a wide range of tasks involving contrastive learning. Thus, developing a more suitable technique for generating hard negatives is a promising direction.

6.2.2 Sampling Bias for Negative Samples. According to debiased contrastive learning [21], using random sampling or remaining in-batch samples as negative samples in self-supervised contrastive learning can result in sampling bias. In other words, the absence of label information results in the intervention of false negatives (samples that have the same label as the anchor but are

regarded as negative samples), which in turn deteriorates the contrastive learning performance. To overcome this issue, Chuang et al. [21] proposed the debiased contrastive loss, which utilizes the data distribution of positive samples and the probability of being chosen as a positive sample to derive the data distribution of negative samples. However, Chen et al. [17] pointed out that such a method cannot explicitly tackle false negatives; they devised an incremental clustering technique to dynamically detect and remove false negatives to eliminate the adverse effects of false negatives.

Although some studies [17, 21, 140] have demonstrated that removing sampling bias for negative samples in self-supervised contrastive learning can result in significant and substantial improvements, only DCLR [142] has identified sampling bias for negative samples in the field of sentence representation learning, and it mitigates the performance degradation caused by sampling bias by assigning a weight of zero to false negatives. DCLR detects false negatives by calculating the cosine similarity between the anchor and negative samples, with SimCSE acting as an encoder to generate sentence representations. However, this method of screening for false negatives necessitates the use of the model SimCSE, which complicates the implementation of the screening method. Prospective work could continue to explore ways of finding false negatives in negative samples.

6.3 Incorporate Effective Datasets

Constructing positive pairs and hard negatives is important for contrastive learning, and experimental results from the work of Chuang et al. [22] and Wang et al. [107] suggest that introducing datasets that are highly similar to the training corpus can help models capture delicate but meaningful semantic features. For example, DiffCSE [22] employs ELECTRA to regenerate edited sentences to improve the model's sensitivity to subtle noise; SNCSE [107] introduces soft negative samples to enable the model to identify differences between anchor points and similar texts with different semantics, thereby improving the stability of the model. Moreover, EASE [76] shows that incorporating external supervision from the Wikipedia hyperlink for entity-aware contrastive learning provides rich training signals for sentence representations. Therefore, future research could strive to construct suitable textual datasets or use annotated textual datasets related to the training corpus so that the model can draw out more intrinsic sentence features.

6.4 Contrastive Loss Function

The design of the loss function has been a crucial component of contrastive learning. The contrastive loss function is defined by measuring the semantic representation distance of input examples in the projection space. Therefore, selecting a suitable distance function is important in contrastive learning. The widely used distance functions in contrastive loss are mainly Manhattan distance (L1-norm), Euclidean distance (L2-norm), and similarity functions (cosine similarity and bilinear similarity). In contrast, non-Euclidean distance functions such as the Jaccard distance and the Hamming distance have not been explored. Thus, researchers should attempt to use these distance functions to measure the similarity between representations. In addition to modifying the distance function, ArcCSE [139] adds a margin constant m to the cosine similarity of the positive pairs, making the model more resistant to noise and enhancing the model performance. Future research could attempt to appropriately modify the InfoNCE loss.

6.5 Multimodal Contrastive Learning

Multimodal learning is gaining popularity as the intra-relationship between multimodal data allows for better exploitation of the intrinsic data properties of each modality. One study [130] showed that employing contrastive learning to train models on multimodal data improved the quality of learned visual representations. Wang et al. [111] also showed that multimodal

contrastive pretraining facilitated the extraction of good code representations for the syntax and structure information provided by multimodal data. Although contrastive learning with sentence-image pairs as the multimodal data has been used for sentence embedding [136], potential research could consider experimenting with sentence-audio pairs as multimodal data to learn sentence representations, or combining multimodal sentence-image-audio data to learn more general semantic features.

6.6 More CV-inspired Models

Numerous contrastive learning based sentence representation models, such as IS-BERT, ConSERT, and DiffCSE, have been inspired by analogous approaches in CV. Therefore, more CV-inspired models are expected to emerge. For instance, negative DA [94] involves generating out-of-distribution samples, which has been explored in CV but not studied in text [31]. The hard negative mining technique "Ring" [116] in CV, which uses a family of MI estimators to sample negatives in a ring around each positive sample, has also not yet been researched in text. Additionally, new contrastive learning models, such as mutual contrastive learning [126] and contrastive continual learning [15] in CV, could be considered for sentence representation learning.

7 CONCLUSION

This article presented a comprehensive and structured study of contrastive learning models for sentence representations. We observed that the success of contrastive learning heavily depends on the selection of the model framework, positive pairs, and negative pairs. In particular, we conducted a series of experiments on evaluation tasks. The experimental results showed that adopting effective DA techniques can be beneficial for transformer-based PLMs in contrastive learning. Both the construction of hard negatives and the incorporation of external training data associated with the Wikipedia training corpus improved the quality of learned representations. Alignment and uniformity analysis demonstrated that the models based on contrastive learning greatly improved the uniformity of pretrained BERT embeddings while maintaining good alignment. The t-SNE visualization of Stack Overflow representations revealed the performance differences between the models. Finally, we outlined the research directions that could be further explored.

REFERENCES

[1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, et al. 2015. SemEval-2015 Task 2: Semantic textual similarity, English, Spanish and Pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 252–263. 10.18653/v1/S15-2045

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'14)*. 81–91. 10.3115/v1/S14-2010

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 497–511. 10.18653/v1/S16-1081

[4] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the Main Conference and the Shared Task ('SEM'12), and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval'12)*. 385–393.

[5] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity ('SEM'13)*. 32–43. https://aclanthology.org/S13-1004.

[6] Holger Schwenk, Loïc Barrault, Alexis Conneau, Douwe Kiela, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*.

[7] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*.

[8] Zhidong Bai and Jianfeng Yao. 2012. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis* 106 (2012), 167–177.

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

[10] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. 632–642. 10.18653/v1/D15-1075

[11] Rui Cao, Yihao Wang, Yuxin Liang, Ling Gao, Jie Zheng, Jie Ren, and Zheng Wang. 2022. Exploring the impact of negative samples of contrastive learning: A case study of sentence embedding. *arXiv preprint arXiv:2202.13093* (2022).

[12] Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*. https://openreview.net/forum?id=Ov_sMNau-PF.

[13] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. 1–14. 10.18653/v1/S17-2001

[14] Daniel Cer, Yinfei Yang, Sheng-Yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, et al. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP'18)*. 169–174. 10.18653/v1/D18-2029

[15] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. 2021. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR'21)*. 9516–9525.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML'20)*. 1597–1607.

[17] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. 2022. Incremental false negative detection for contrastive learning. In *Proceedings of the International Conference on Learning Representations (ICLR'22)*. https://openreview.net/forum?id=dDjSKKA5TP1.

[18] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 15750–15758.

[19] Xiang Chen, Xin Xie, Zhen Bi, Hongbin Ye, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. Disentangled contrastive learning for learning robust textual representations. In *Proceedings of the CAAI International Conference on Artificial Intelligence*. 215–226.

[20] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, Los Alamitos, CA, 539–546.

[21] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 8765–8775.

[22] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-Tau Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. *arXiv preprint arXiv:2204.10298* (2022).

[23] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[24] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the International Conference on Learning Representations (ICLR'20)*. https://openreview.net/pdf?id=r1xMH1BtvB.

[25] Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. https://aclanthology.org/L18-1269.

[26] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. 2022. Equivariant self-supervised learning: Encouraging equivariance in representations. In *Proceedings of the International Conference on Learning Representations (ICLR'22)*. https://openreview.net/forum?id=gKLAAfiytI.

[27] 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)*. 4171–4186. 10.18653/v1/N19-1423

[28] William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP'05)*. https://aclanthology.org/I05-5002.

[29] Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. 55–65. 10.18653/v1/D19-1006

[30] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766* (2020).

[31] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 968–988. 10.18653/v1/2021.findings-acl.84

[32] Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, Vol. 7. 1606–1611.

[33] Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He, and Lawrence Carin. 2016. Unsupervised learning of sentence representations using convolutional neural networks. *arXiv preprint arXiv:1611.07897* (2016).

[34] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*. https://openreview.net/forum?id=SkEYojRqtm.

[35] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 6894–6910. 10.18653/v1/2021.emnlp-main.552

[36] John Giorgi, Osvald Nitski, Bo Wang, and Gary 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 (Volume 1: Long Papers)*. 879–895. 10.18653/v1/2021.acl-long.72

[37] Fausto Giunchiglia and Pavel Shvaiko. 2003. Semantic matching. *Knowledge Engineering Review* 18, 3 (2003), 265–280.

[38] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, et al. 2020. Bootstrap your own latent—A new approach to self-supervised learning. *Advances in Neural Information Processing Systems* 33 (2020), 21271–21284.

[39] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. 297–304.

[40] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, Los Alamitos, CA, 1735–1742.

[41] 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 (CVPR'20)*. 9729–9738.

[42] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483* (2016).

[43] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[44] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).

[45] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. ACM, New York, NY, 168–177. 10.1145/1014052.1014073

[46] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. 2021. AdCo: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 1074–1083.

[47] Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. 2022. The power of contrast for feature learning: A theoretical analysis. In *Proceedings of the International Conference on Learning Representations (ICLR'22)*. https://openreview.net/pdf?id=yBYVUDj7yF.

[48] Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337* (2022).

[49] Yuxin Jiang and Wei Wang. 2022. Deep continuous prompt for contrastive learning of sentence embeddings. *arXiv preprint arXiv:2203.06875* (2022).

[50] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 21798–21809.

[51] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* (2014).

[52] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[53] Hwi-Gang Kim, Seongjoo Lee, and Sunghyon Kyeong. 2013. Discovering hot topics using Twitter streaming data social topic detection and geographic clustering. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'13)*. IEEE, Los Alamitos, CA, 1215–1220.

[54] Taeuk Kim, Kang Min Yoo, and Sang-Goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2528–2540. 10.18653/v1/2021.acl-long.197

[55] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1746–1751. 10.3115/v1/D14-1181

[56] Ryan Kiros, Yukun Zhu, Russ R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in Neural Information Processing Systems* 28 (2015), 3294–3302.

[57] Haejun Lee, Drew A. Hudson, Kangwook Lee, and Christopher D. Manning. 2020. SLM: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. 1551–1562. 10.18653/v1/2020.emnlp-main.120

[58] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. 9119–9130. 10.18653/v1/2020.emnlp-main.733

[59] Renhao Li, Lei Duan, Guicai Xie, Shan Xiao, and Weipeng Jiang. 2022. AdCSE: An adversarial method for contrastive learning of sentence embeddings. In *Database Systems for Advanced Applications*, Arnab Bhattacharya, Janice Lee Mong Li, Divyakant Agrawal, P. Krishna Reddy, Mukesh Mohania, Anirban Mondal, Vikram Goyal, and Rage Uday Kiran (Eds.). Springer International Publishing, Cham, Switzerland, 165–180.

[60] Xianming Li, Zongxi Li, Haoran Xie, and Qing Li. 2021. Merging statistical feature via adaptive gate for improved text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13288–13296.

[61] Xianming Li, Zongxi Li, Yingbin Zhao, Haoran Xie, and Qing Li. 2020. Incorporating effective global information via adaptive gate attention for text classification. *arXiv preprint arXiv:2002.09673* (2020).

[62] Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning cross-context entity representations from text. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepLANLP'21)*. 241–247. https://openreview.net/forum?id=HygwvC4tPH.

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

[64] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137.

[65] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*. https://openreview.net/forum?id=rJvJXZb0W.

[66] Wenpeng Lu, Xu Zhang, Huimin Lu, and Fangfang Li. 2020. Deep hierarchical encoding model for sentence semantic matching. *Journal of Visual Communication and Image Representation* 71 (2020), 102794.

[67] J. MacQueen. 1967. Classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. 281–297.

[68] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the 9th*

International Conference on Language Resources and Evaluation (LREC'14). 216–223. http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

[69] Michael Mathioudakis and Nick Koudas. 2010. TwitterMonitor: Trend detection over the Twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. 1155–1158.

[70] Deshui Miao, Jiaqi Zhang, Wenbo Xie, Jian Song, Xin Li, Lijuan Jia, and Ning Guo. 2021. Simple contrastive representation adversarial learning for NLP tasks. *arXiv preprint arXiv:2111.13301* (2021).

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

[72] George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38, 11 (1995), 39–41.

[73] Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2018), 1979–1993.

[74] Ipsita Mohanty, Ankit Goyal, and Alex Dotterweich. 2021. Emotions are subtle: Learning sentiment based text representations using contrastive learning. *arXiv preprint arXiv:2112.01054* (2021).

[75] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 1864–1874. 10.18653/v1/2022.findings-acl.146

[76] Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-aware contrastive learning of sentence embedding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'22)*.

[77] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 4004–4012.

[78] Bruno A. Olshausen and David J. Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37, 23 (1997), 3311–3325.

[79] Bruno A. Olshausen and David J. Field. 2004. Sparse coding of sensory inputs. *Current Opinion in Neurobiology* 14, 4 (2004), 481–487.

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

[81] Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 244–258. 10.18653/v1/2021.acl-long.21

[82] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*. 271–278. 10.3115/1218955.1218990

[83] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 115–124. 10.3115/1219840.1219855

[84] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543. 10.3115/v1/D14-1162

[85] 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*. 91–100.

[86] Md. Rashadul Hasan Rakib, Norbert Zeh, Magdalena Jankowska, and Evangelos Milios. 2020. Enhancement of short text clustering by iterative classification. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems*. 105–117.

[87] Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING'16)*. 87–96.

[88] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. 3982–3992. 10.18653/v1/D19-1410

[89] Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400* (2018).

[90] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). http://arxiv.org/abs/1910.01108.

[91] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 815–823.

[92] Hooman Sedghamiz, Shivam Raval, Enrico Santus, Tuka Alhanai, and Mohammad Ghassemi. 2021. SupCL-Seq: Supervised contrastive learning for downstream optimized sequence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 3398–3403. 10.18653/v1/2021.emnlp-main.289

[93] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818* (2020).

[94] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. 2021. Negative data augmentation. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*. https://openreview.net/forum?id=Ovp8dvB8IBH.

[95] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*. 1631–1642. https://aclanthology.org/D13-1170.

[96] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class N-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. 1857–1865.

[97] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[98] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316* (2021).

[99] Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 4381–4394. 10.18653/v1/2021.emnlp-main.359

[100] Haochen Tan, Wei Shao, Han Wu, Ke Yang, and Linqi Song. 2022. A sentence is worth 128 pseudo tokens: A semantic-aware contrastive learning framework for sentence embeddings. *arXiv preprint arXiv:2203.05877* (2022).

[101] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision*. 776–794.

[102] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 1–27.

[103] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS'17)*.

[104] Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 200–207.

[105] Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: Contrastive learning with semantic negative examples for natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2332–2342. 10.18653/v1/2021.acl-long.181

[106] Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 2495–2504.

[107] Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. SNCSE: Contrastive learning for unsupervised sentence embedding with soft negative samples. *CoRR abs/2201.05979* (2022).

[108] Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 671–688. 10.18653/v1/2021.findings-emnlp.59

[109] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML'20)*. 9929–9939.

[110] Wei Wang, Liangzhu Ge, Jingqiao Zhang, and Cheng Yang. 2022. Improving contrastive learning of sentence embeddings with case-augmented positives and retrieved negatives. *arXiv preprint arXiv:2206.02457* (2022).

[111] Xin Wang, Yasheng Wang, Fei Mi, Pingyi Zhou, Yao Wan, Xiao Liu, Li Li, Hao Wu, Jin Liu, and Xin Jiang. 2021. Syncobert: Syntax-guided multi-modal contrastive pre-training for code representation. *arXiv preprint arXiv:2108.04556* (2021).

[112] Zixin Wen and Yuanzhi Li. 2021. Toward understanding the feature learning process of self-supervised contrastive learning. In *Proceedings of the International Conference on Machine Learning (ICML'21)*. 11112–11122.

[113] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39, 2 (2005), 165–210.

[114] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1112–1122. 10.18653/v1/N18-1101

[115] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45. 10.18653/v1/2020.emnlp-demos.6

[116] Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, and Noah Goodman. 2021. Conditional negative sampling for contrastive learning of visual representations. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*. https://openreview.net/forum?id=v8b3e5jN66j.

[117] Xing Wu, Chaochen Gao, Jue Wang, Liangjun Zang, Zhongyuan Wang, and Songlin Hu. 2021. DisCo: Effective knowledge distillation for contrastive learning of sentence embeddings. *arXiv preprint arXiv:2112.05638* (2021).

[118] Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics*. 3898–3907. https://aclanthology.org/2022.coling-1.342.

[119] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466* (2020).

[120] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*. https://openreview.net/forum?id=zeFrfgyZln.

[121] 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 (2017), 22–31.

[122] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. 2020. Hard negative examples are hard, but useful. In *Proceedings of the European Conference on Computer Vision*. 126–142.

[123] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning distributed representations of texts and entities from knowledge base. *Transactions of the Association for Computational Linguistics* 5 (2017), 397–411.

[124] Ikuya Yamada, Hiroyuki Shindo, and Yoshiyasu Takefuji. 2018. Representation learning of entities and documents from knowledge base descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*. 190–201. https://aclanthology.org/C18-1016.

[125] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

[126] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. 2022. Mutual contrastive learning for visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3045–3053.

[127] Ziyi Yang, Yinfei Yang, Daniel Cer, Law Jax, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 6216–6228. 10.18653/v1/2021.emnlp-main.502

[128] Jianfeng Yao, Shurong Zheng, and Z. D. Bai. 2015. *Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press.

[129] Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering (ICDE'16)*. IEEE, Los Alamitos, CA, 625–636.

[130] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 6995–7004.

[131] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the International Conference on Machine Learning (ICML'21)*. 12310–12320.

[132] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1503–1512.

[133] Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 5786–5798. 10.18653/v1/2021.emnlp-main.467

[134] Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew O. Arnold. 2021. Virtual augmentation supported contrastive learning of sentence representations. *arXiv preprint arXiv:2110.08552* (2021).

[135] Han Zhang, Zongxi Li, Haoran Xie, Raymond YK Lau, Gary Cheng, Qing Li, and Dian Zhang. 2022. Leveraging statistical information in fine-grained financial sentiment analysis. *World Wide Web* 25, 2 (2022), 513–531.

[136] Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A. Hedderich, and Dietrich Klakow. 2022. MCSE: Multimodal contrastive learning of sentence embeddings. *arXiv preprint arXiv:2204.10931* (2022).

[137] 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)*. 1601–1610. 10.18653/v1/2020.emnlp-main.124

[138] Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22)*.

[139] Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4892–4903. 10.18653/v1/2022.acl-long.336

[140] Han Zhao, Xu Yang, Zhenru Wang, Erkun Yang, and Cheng Deng. 2021. Graph debiased contrastive learning with joint representation clustering. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*. 3434–3440.

[141] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. 2019. Hardness-aware deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 72–81.

[142] Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. *arXiv preprint arXiv:2205.00656* (2022).

Received 6 September 2022; revised 15 March 2023; accepted 12 April 2023