Zero-Shot Stance Detection via Contrastive Learning

Bin Liang[*]
Harbin Institute of Technology
Joint Lab of CMS-HITSZ
Shenzhen, China
bin.liang@stu.hit.edu.cn

Zixiao Chen[*]
Harbin Institute of Technology
Joint Lab of CMS-HITSZ
Shenzhen, China
chenzixiao@stu.hit.edu.cn

Lin Gui
University of Warwick
Coventry, UK
lin.gui@warwick.ac.uk

Yulan He
University of Warwick
Coventry, UK
Yulan.He@warwick.ac.uk

Min Yang
SIAT, Chinese Academy of Sciences
Shenzhen, China
min.yang@siat.ac.cn

Ruifeng Xu[†]
Harbin Institute of Technology
Peng Cheng Lab
Shenzhen, China
xuruifeng@hit.edu.cn

ABSTRACT

Zero-shot stance detection (ZSSD) is challenging as it requires detecting the stance of previously unseen targets during the inference stage. Being able to detect the target-related transferable stance features from the training data is arguably an important step in ZSSD. Generally speaking, stance features can be grouped into target-invariant and target-specific categories. Target-invariant stance features carry the same stance regardless of the targets they are associated with. On the contrary, target-specific stance features only co-occur with certain targets. As such, it is important to distinguish these two types of stance features when learning stance features of unseen targets. To this end, in this paper, we revisit ZSSD from a novel perspective by developing an effective approach to distinguish the types (target-invariant/-specific) of stance features, so as to better learn transferable stance features. To be specific, inspired by self-supervised learning, we frame the stance-feature-type identification as a pretext task in ZSSD. Furthermore, we devise a novel hierarchical contrastive learning strategy to capture the correlation and difference between target-invariant and -specific features and further among different stance labels. This essentially allows the model to exploit transferable stance features more effectively for representing the stance of previously unseen targets. Extensive experiments on three benchmark datasets show that the proposed framework achieves the state-of-the-art performance in ZSSD.

CCS CONCEPTS

• **Information systems** → *Sentiment analysis.*

KEYWORDS

zero-shot stance detection, contrastive learning, pretext task

[*]The first two authors contribute equally to this work.
[†]Corresponding Author.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3511994

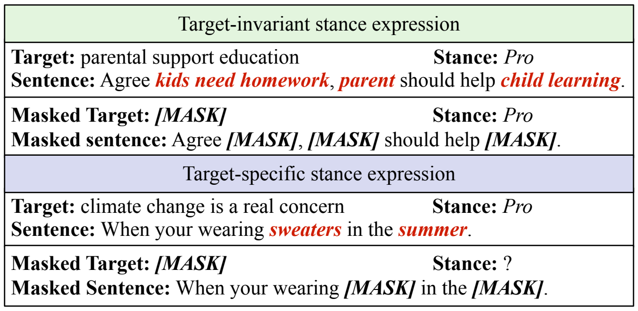
ACM Reference Format:
Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. Zero-Shot Stance Detection via Contrastive Learning. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3511994

1 INTRODUCTION

The aim of stance detection is to determine people's opinionated standpoint or attitude (e.g. *Favor*, *Against*, or *Neutral*, etc.) expressed in text towards a specific target, topic or proposition [3, 27, 35], which is a vital task in natural language processing (NLP) [18]. In traditional in-target stance detection [16], the same set of targets are assumed to be seen in both the training and the test sets. However, in real world scenarios, the targets during inference might be unseen/unknown to a trained stance detection model [1], since it is infeasible to enumerate all possible targets beforehand during model training. Cross-target stance detection [3, 43] partially alleviates this issue by adapting classifiers trained on a certain target to a related new target. But this is based on a strong assumption that knowing the target relation in advance. To address the problem of target mismatch between training and testing, we proposed a zero-shot stance detection (ZSSD) framework, which aims to perform stance detection on unseen targets, a more realistic setup.

To deal with ZSSD, some existing methods attempt to introduce attention mechanisms [1] or external knowledge to capture relationships between targets [24], in order to generalize to unseen targets. However, in practice, directly transferring stance features from seen targets to unseen ones may not lead to good results since there might be features specific to certain targets. As such, it is important to distinguish target-invariant features, which carry the same stance regardless of the targets they are associated with, from target-specific ones, which only co-occur with certain targets.

To illustrate our idea, we give examples of target-invariant and -specific stance expressions in Figure 1. Words that are indicative of a target are highlighted in red italic. In the target-invariant stance expression, the original stance expressed in the context can still be identified even when the target or target-related words are masked. While in the target-specific one, it is difficult to make sense of the stance information if the target and the target-related words are masked. That is, these two types of stance features play different roles in learning stance information for the unseen targets.



Target-invariant stance expression

Target: parental support education **Stance:** *Pro*
Sentence: Agree ***kids need homework***, ***parent*** should help ***child learning***.

Masked Target: ***[MASK]*** **Stance:** *Pro*
Masked sentence: Agree ***[MASK]***, ***[MASK]*** should help ***[MASK]***.

Target-specific stance expression

Target: climate change is a real concern **Stance:** *Pro*
Sentence: When your wearing ***sweaters*** in the ***summer***.

Masked Target: ***[MASK]*** **Stance:** ?
Masked Sentence: When your wearing ***[MASK]*** in the ***[MASK]***.

Figure 1: Examples of the types of stance expressions.

We hence propose to study how to automatically detect the types (target-invariant/-specific) of stance features in the training data. To identify the target-invariant stance features, some existing research efforts focus on utilizing adversarial learning by introducing a discriminator to deal with an auxiliary task of target classification [2, 41]. However, in practice, the imbalanced distribution of stance targets may limit the performance of the discriminator trained on labeled targets. We propose an alternative unsupervised approach which distinguishes target-invariant and target-specific stance features by examining the stance change when masking a target and the target-related words.

More concretely, for each target, we first derive a set of target-related words and determine their relatedness weights. Here, the formulation of deriving target-related words could be based on relatedness measures, including but not limited to TF-IDF, word similarity, and topic modeling results. After that, the target-related words will be selected as candidates to be masked according to their weights in order to generate masked training instances. Then, we explore a simple but effective solution that feeds the masked instances into a well-trained stance detection model (training accuracy is close to 100%) to predict their stance labels. The prediction results, subsequently, are exploited to generate the surrogate supervision signal of pretext task for contrastive self-supervised learning. That is, intuitively, the stance expression is target-invariant if the predicted label of the masked instance is correct, otherwise it is target-specific. Finally, inspired by [5], to improve the quality of the learned embeddings by distinguishing the types of stance features in the latent distribution space, we devise a novel hierarchical contrastive learning strategy that considers both the surrogate supervision signal and the stance label information. We call the proposed framework Pretext Task-based Hierarchical Contrastive Learning (**PT-HCL**). This enables the transferable stance features to be leveraged for understanding the stance expressions of unseen targets, and thus leads to improved ZSSD performance. The main contributions of our work can be summarized as follows:

- The ZSSD task is approached from a new perspective that the types (target-invariant/-specific) of stance features are distinguished in the latent distribution space, so as to preferably leverage the target-oriented transferable stance features for stance detection of previously unseen targets.
- A novel scenario of deriving surrogate supervised signals for self-supervised feature learning is explored by predicting the masked variants of the well-trained instances, in which the training sentences are masked in the light of the target-related words to generate masked instances.
- Based on the surrogate supervised signal of the pretext task, we devise a novel hierarchical contrastive learning framework (PT-HCL) to improve the quality of learned representations by considering both the types of stance features and the stance labels, enabling preferably generalize the learning ability of the model to deal with ZSSD.
- Extensive experiments on 3 benchmark datasets show that the proposed framework achieves state-of-the-art performance in ZSSD. We also extend the proposed framework to few-shot and cross-target stance detection to demonstrate the superiority and generalization of our approach.

2 RELATED WORK

2.1 Zero-shot Stance Detection

Previous studies of stance detection largely focus on target-specific stance detection, where the training and inference stages share the same pre-defined set of targets [3, 11, 19, 22, 33, 36]. In previous research, a task similar to ZSSD is the cross-target stance detection, where the learning ability of the classifier is adaptive to the unseen but related targets in the light of training on a known one [23, 41, 43, 45]. Existing cross-target stance detection research efforts generally explored attention-based models [41, 43] or graph-based models [23, 45] to learn a projection that aims to transfer the target-related stance features from a specific training target to adapt the related testing one. Different from cross-target stance detection, zero-shot stance detection (ZSSD) aims to automatically identify the stance of the previously unseen targets, which is a more accurate evaluation of a model's ability to generalize to the newly emerging targets in the real world scenario [1]. To deal with zero-shot stance detection, [8] presented a large-scale expert-annotated Twitter stance detection dataset, where the testing targets are unknown to the training targets set. Allaway and McKeown [1] created a ZSSD dataset consists of a large range of topics covering broad themes and proposed a topic-grouped attention model to implicitly capture relationships between targets by using generalized topic representations. [2] adapted a target-specific stance detection dataset [27] to ZSSD, and deployed adversarial learning to extract target-invariant transformation features in ZSSD. Further, to exploit both the structural-level and semantic-level information of the relational knowledge, [24] proposed a commonsense knowledge enhanced graph model based on BERT [9] to cope with ZSSD.

2.2 Pretext Task and Contrastive Learning

Self-supervised learning is a prominent research paradigm where the supervisory signal for feature learning is automatically generated from the data itself. The recent renaissance of self-supervised learning began with artificially designed prediction tasks, often referred to as pretext tasks [5, 12, 47]. Many existing computer vision approaches have designed annotation free pretext tasks based on heuristics such that providing a surrogate supervision signal of feature learning for the target problems [12, 21, 34, 47]. Such as relative patch prediction [10], solving jigsaw puzzles [28], and image rotation prediction [7, 12], etc. Further, contrastive learning in the latent space has recently shown great promise in self-supervised learning, allowing the automatically generated supervisory signal to be effectively close the gap with fully-supervised learning and improve the quality of the learned representations of the training

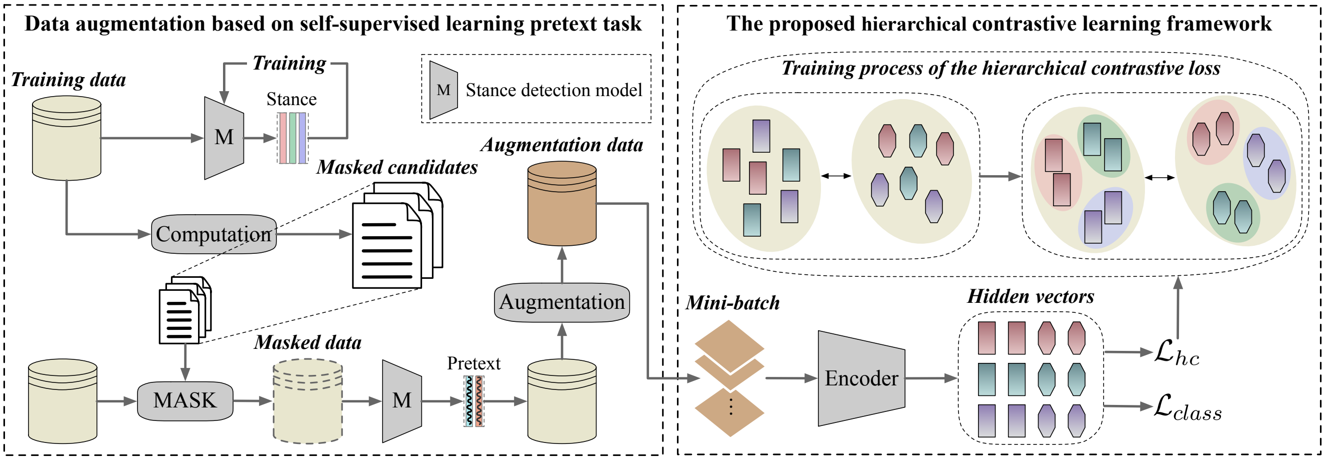


Figure 2: The architecture of the proposed PT-HCL framework. Shapes in gradient colors represent hidden vectors, and different shapes represent different labels of the surrogate supervised signal provided by the pretext task.

data [5, 6, 15, 20, 26, 37, 42]. On the other hand, recent literature has attempted to design their methods by relating contrastive learning for natural language processing [17, 25, 30, 44], such as text clustering [46], unsupervised textual representations [13], text classification [31], and multilingual neural machine translation [29], etc. As such, inspired by existing remarkable self-supervised learning approaches, we attempt to deploy a novel Pretext Task-based Hierarchical Contrastive Learning (PT-HCL) framework to preferably generalize the target-oriented transferable stance features learned from known targets to the unseen targets for dealing with ZSSD.

3 METHODOLOGY

In this section, we describe our proposed **PT-HCL** framework for zero-shot stance detection in detail. As demonstrated in Figure 2, the proposed PT-HCL framework mainly consists of two modules: *data augmentation based on the self-supervised learning pretext task* (**data augmentation via pretext task**) and *the proposed contrastive learning framework* (**hierarchical contrastive learning framework**). The first module, **data augmentation via pretext task**, mainly contains three steps: 1) training an over-fitting stance detection model $\mathcal{M}$; 2) deriving masked candidates for each target; and 3) generating augmentation data for the training set. The module, **hierarchical contrastive learning framework**, mainly contains three components: 1) the encoder; 2) hierarchical contrastive learning; and 3) the stance classifier.

3.1 Task Description

Formally, supposing there is a set of annotated instances towards source targets $\mathcal{D}_s = \{(r_s^i, t_s^i, y_s^i)\}_{i=1}^{N_s}$ and a set of unlabeled instances towards unseen or testing targets $\mathcal{D}_d = \{(r_d^i, t_d^i)\}_{i=1}^{N_d}$, where y_s^i is the stance label of an annotated instance towards the source target t_s, N_s and N_d are the number of instances towards the training and testing targets, respectively. Note that there is no overlap between $\mathcal{D}_s$ and $\mathcal{D}_d$. The goal of zero-shot stance detection is to train a model from each sentence r_s^i towards the source known target t_s^i from $\mathcal{D}_s$, which can be generalized to detect stance for each sentence r_d^i towards an unseen target t_d^i in $\mathcal{D}_d$.

Table 1: Examples of topic words.

Target	Topic words
parental support education	learn, kid, need, problem, homework, paren, child, learning, teach, book
American unemployment	unemployment, americans, neighbors, paid, job, pay, cost, outsourcing, communities, rate

3.2 Data Augmentation via Pretext Task

To distinguish the types (target-invariant/-specific) of stance features in order to better learn transferable stance features for the unseen targets in ZSSD, we explore a novel strategy of data augmentation based on a target-aware self-supervised learning pretext task and a well-trained stance detection model $\mathcal{M}$. Here, inspired by existing methods [5, 12, 34, 47], we design the task of distinguishing the types of stance features as a pretext task in order to provide a surrogate supervised signal for contrastive learning.

3.2.1 Training a Vanilla Stance Detection Model from the Training Data. To learn a good representation for each training instance in $\mathcal{D}_s$, we first train a vanilla stance detection model $\mathcal{M}$ with $\mathcal{D}_s$. For each instance consisting of a sentence r_s^i and a target t_s^i, we employ the pre-trained uncased BERT-base [9] as the stance detection model $\mathcal{M}$, which takes "$[CLS] r_s^i [SEP] t_s^i [SEP]$" as input. We use the vector of token $[CLS]$ to represent the input instance for stance detection. Here, the training accuracy of $\mathcal{M}$ is close to 100%.

3.2.2 Deriving Masked Candidates for Each Target. The aim of deriving masked candidates for each target is to extract a set of target-related words and derive their relatedness weights, which can be later selected as candidates to be masked according to their weights. Methods used to identify related words include TF-IDF, word similarity measures, and topic models. In this work, we deploy latent Dirichlet allocation (LDA) [4] to capture topic words for each target[1]. Following the LDA implementation of [14], the documents of each target could be expressed by T topics and each topic contains K words. As such, the corresponding topic words can be regarded as a bridge to leverage the target-related words in the context, allowing the crucial clues of the mentioned target to be explicitly used in stance features learning. Table 1 presents the top 10 topic words of two example targets. We observe that the topic words are quite relevant to the corresponding target.

3.2.3 Generating Augmentation Data. To distinguish the target-invariant and target-specific stance features for better learning transferable stance features in ZSSD, we design a pretext task via self-supervised learning to automatically generate auxiliary supervisory signals from the training data. Specifically, as the examples shown in Figure 1, we mask the topic words in each training sentence with a special token $[MASK]$. Then, we feed the masked sentence paired with the masked target $[MASK]$ into the vanilla stance detection model $\mathcal{M}$ to predict the stance label of this instance. If the predicted label is correct, which implies its stance expression is not dependent on the target, then the stance expression of the instance is target-invariant. Accordingly, we attach an augmentation label of "*target-invariant*" to this instance. Otherwise, the augmentation label is "*target-specific*". Formally, after data augmentation, the training set can be represented as $\mathcal{D}_s = \{(r^i_s, t^i_s, y^i_s, p^i_s)\}^{N_s}_{i=1}$.

3.3 The Proposed Framework

3.3.1 Encoder Module. Given a sequence of words $r = \{w_i\}^n_{i=1}$ and the corresponding target t, n is the length of the text r. Here, we use r and t to represent the sentence and the target of a training instance. Then, we adopt a pre-trained BERT [9] as the Encoder Module and feed "$[CLS]r[SEP]t[SEP]$" as input into the encoder module to acquire a d_m-dimensional hidden representation $\boldsymbol{h} \in \mathbb{R}^{d_m}$ of the token $[CLS]$ of each input sample:

$$\boldsymbol{H} = \mathrm{BERT}([CLS]r[SEP]t[SEP]), \boldsymbol{h} = \boldsymbol{H}_{[CLS]} \tag{1}$$

That is, for a mini-batch, the hidden representations of the samples can be defined as: $\mathcal{B} = \{\boldsymbol{h}_i\}^{N_b}_{i=1}$, N_b is the size of mini-batch.

3.3.2 Hierarchical Contrastive Learning. Contrastive learning is designed to learn with a pair-based contrastive loss function, allowing the representation of a given anchor data to be similar to the *positive* data and dissimilar to the *negative* data, such as the supervised contrastive learning loss proposed by Khosla et al. [20] and the NT-Xent loss used in [5]. For each mini-batch $\mathcal{B}$, the general contrastive loss function is formulated as:

$$\ell_{i,j} = -\log \frac{\exp(z_i, z_j/\tau)}{\sum^{N_b}_{k=1} \mathbb{1}_{[k \neq i]} \exp(z_i, z_k/\tau)} \tag{2}$$

where z_i is the anchor, $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$, τ denotes a temperature parameter.

Further, Wang and Liu [39] has demonstrated that the hardness-aware property is significant to the success of contrastive loss, and the temperature parameter τ is a key factor to control the strength of penalties on hard negative samples in contrastive learning. Contrastive loss with *small temperature* tends to penalize much more on the hardest negative samples such that the local structure of each sample tends to be more separated, and the embedding distribution is likely to be more uniform.

Motivated by these, we devise a novel hierarchical contrastive learning loss to take into account both the types of stance features and the information of stance labels. We first make contrastive representations between target-invariant and target-specific stance features in the latent distribution space. Based on it, for the latent space of each stance feature role, we further perform contrastive learning among the stance labels. This allows the contrastive representations of stance information to be thought over when preferentially performing the contrastive learning of stance feature types. More concretely, we explore two temperature parameters with different values regarding the hierarchical contrastive loss. This aims to make the aggregation and separation in the latent space of the stance features types sharply clearer via a *smaller* temperature parameter τ, and simultaneously grasping the correlation and difference among different stance labels without disturbing the latent space of the types via a *larger* temperature parameter τ.

To be specific, for an *anchor* $\boldsymbol{h}_i$, we refer to $\boldsymbol{h}_i, \boldsymbol{h}_j \in \mathcal{B}$ with the same augmentation label of pretext task (the role of stance features) as a *primary positive* pair, i.e. $p^i = p^j$, and with the same stance polarity as a *secondary positive* pair, i.e. $y^i = y^j$. The samples $\{\boldsymbol{h}_k \in \mathcal{B}, k \neq i\}$ are treated as *negative* instances regarding this *anchor*. Note that, the contrastive loss is computed across all *positive* pairs, both $(\boldsymbol{h}_i, \boldsymbol{h}_j)$ and $(\boldsymbol{h}_j, \boldsymbol{h}_i)$ in a mini-batch. Then, we extend the contrastive loss of Eq. 2 and define the hierarchical contrastive loss of each mini-batch $\mathcal{B}$ as:

$$\mathcal{L}_{hc} = \frac{-1}{N_b} \sum_{\boldsymbol{h}_i \in \mathcal{B}} \ell(\boldsymbol{h}_i) \tag{3}$$

$$\begin{aligned} \ell(\boldsymbol{h}_i) = \log\Big(& \frac{\sum^{N_b}_{j=1} \mathbb{1}_{[j \neq i]} \mathbb{1}_{[p^i = p^j]} f(\boldsymbol{h}_i, \boldsymbol{h}_j)}{\sum^{N_b}_{k=1} \mathbb{1}_{[k \neq i]} f(\boldsymbol{h}_i, \boldsymbol{h}_k)} \\ & \times \alpha \frac{\sum^{N_b}_{j=1} \mathbb{1}_{[j \neq i]} \mathbb{1}_{[p^i = p^j]} \mathbb{1}_{[y^i = y^j]} g(\boldsymbol{h}_i, \boldsymbol{h}_j)}{\sum^{N_b}_{k=1} \mathbb{1}_{[k \neq i]} \mathbb{1}_{[p^i = p^k]} g(\boldsymbol{h}_i, \boldsymbol{h}_k)} \Big) \end{aligned} \tag{4}$$

$$f(\boldsymbol{u}, \boldsymbol{v}) = \exp(sim(\boldsymbol{u}, \boldsymbol{v})/\tau_p) \tag{5}$$

$$g(\boldsymbol{u}, \boldsymbol{v}) = \exp(sim(\boldsymbol{u}, \boldsymbol{v})/\tau_y) \tag{6}$$

where $\mathbb{1}_{[i=j]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $i = j$. $sim(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$ denotes the cosine similarity between L_2 normalized vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. α, τ_p, and τ_y are tuned hyperparameters to control the separable strength between positive and negative samples from augmentation labels and stance labels. Here, $\tau_p < \tau_y$. In this way, the novel contrastive learning scenario preferentially pulls together the clusters of points belonging to the same augmentation label in embedding space, and further slightly pulls together the clusters of points belonging to the same stance polarity in each separated embedding space. Simultaneously, pushing apart clusters of samples from different augmentation labels.

3.3.3 Stance Classifier. We first feed the hidden vectors of the mini-batch $\mathcal{B} = \{\boldsymbol{h}_i\}^{N_b}_{i=1}$ into a classifier with a softmax function to produce the predicted stance distribution:

$$\hat{\boldsymbol{y}}_i = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{b}) \tag{7}$$

[1] In Section 5.1, we analyze the results using different methods for identifying target-related words, and find no significant performance difference among them. Nevertheless, using topic model achieves slightly better results.

Table 2: Statistics of Vast dataset.

	Train	Dev	Test
# Examples	13477	2062	3006
# Unique Comments	1845	682	786
# Zero-shot Topics	4003	383	600
# Few-shot Topics	638	114	159

Table 3: Statistics of Sem16 and Wtwt datasets.

Dataset	Target	Favor	Against	Neutral	Unrelated
Sem16	DT	148	299	260	-
	HC	163	565	256	-
	FM	268	511	170	-
	LA	167	544	222	-
	A	124	464	145	-
	CC	335	26	203	-
Wt-wt	CA	2469	518	5520	3115
	CE	773	253	947	554
	AC	970	1969	3098	5007
	AH	1038	1106	2804	2949

where $\hat{\boldsymbol{y}}_i \in \mathbb{R}^{d_y}$ is the predicted stance probability of the input instance x_i, d_y is the dimensionality of stance labels. $\boldsymbol{W} \in \mathbb{R}^{d_y \times d_m}$ and $\boldsymbol{b} \in \mathbb{R}^{d_y}$ are trainable parameters.

Based on the predicted stance probability, we adopt a cross-entropy loss between predicted and ground-truth distribution $\boldsymbol{y}_i$ of instance x_i to train the classifier:

$$\mathcal{L}_{class} = -\sum_{i=1}^{N_b}\sum_{j=1}^{d_p} y_i^j \log \hat{y}_i^j \tag{8}$$

3.4 Training and Inference

3.4.1 Training. The learning objective of our proposed model is to train the model by jointly optimizing a supervised loss of stance detection $\mathcal{L}_{class}$ with a contrastive loss of pretext task $\mathcal{L}_{hc}$. The overall loss $\mathcal{L}$ is formulated by summing up three losses together:

$$\mathcal{L} = \gamma_c \mathcal{L}_{class} + \gamma_h \mathcal{L}_{hc} + \lambda ||\Theta||^2 \tag{9}$$

γ_c and γ_h are tuned hyper-parameters. Θ denotes all trainable parameters of the model, λ represents the coefficient of L_2-regularization.

3.4.2 Inference. For a testing example, we first conduct Eq. 1 to derive the hidden representation $\boldsymbol{h}_t$ of the example. Further, the hidden representation $\boldsymbol{h}_t$ is fed into the well-trained **PT-HCL** framework and obtain the predicted stance distribution $\hat{\boldsymbol{y}}_t$ towards $\boldsymbol{h}_t$ through Eq. 7. Finally, we utilize argmax$(\cdot)$ function to output the label index o_t of the example: $o_t = \text{argmax}(\hat{\boldsymbol{y}}_t)$.

4 EXPERIMENTAL SETUP

4.1 Experimental Data

We conduct the experiments on the following three zero-shot stance detection (ZSSD) datasets:

Vast [1]. This dataset contains a large amount of variable topics (targets). Each instance consists of a sentence r, a target t, and a stance polarity y ("*Pro*", "*Con*", or "*Neutral*") towards t. Following [1], we also conduct experiments over few-shot condition, where the development and test sets consist of very few training targets. The statistics of dataset are shown in Table 2.

Sem16 [27]. This dataset contains 6 pre-defined targets across multiple domains. Including *Donald Trump* (DT), *Hillary Clinton* (HC), *Feminist Movement* (FM), *Legalization of Abortion* (LA), *Atheism* (A), and *Climate Change* (CC). Each instance could be classified as *Favor*, *Against* or *Neutral*. Following [2], we regard a target as the zero-shot testing target while train on the other five, and randomly select 15% of the training set as the development data to tune the hyper-parameters. The statistics of dataset are shown in Table 3.

Wt-wt [8]. This dataset contains 4 targets in discussing mergers and acquisition operations between companies. Including *CVS_AET* (CA), *CI_ESRX* (CE), *ANTM_CI* (AC), and *AET_HUM* (AH). Each instance refers to a stance label of *Support* (corresponding to *Favor*), *Refute* (corresponding to *Against*), *Comment* (corresponding to *Neutral*), or *Unrelated*. Following [8], we regard each target as the zero-shot testing target while training on the other three. We also randomly select 15% of the training set as the development data. The statistics of dataset are shown in Table 3.

4.2 Experimental Implementation

4.2.1 Training Settings. We use the pre-trained uncased BERT-base [9] as the encoder with 768-dimensional embedding, and the learning rate is $5e^{-6}$. Following [43], the coefficient of L_2-regularization λ is set to $1e^{-5}$. Adam is utilized as the optimizer. The mini-batch is set to 16. For contrastive loss, we set the hyper-parameters $\tau_p = 0.07$, $\tau_y = 0.14$, $\alpha = 0.5$, $\gamma_c = 0.8$ and $\gamma_h = 1$. They are the optimal hyper-parameters in the pilot studies[2]. We use latent Dirichlet allocation (LDA) [4] to generate topic words for each target. According to the average data amount of targets, we set $T = 1$ for Vast dataset, and set $T = 10$ for Sem16 and Wt-wt datasets. We select the top $K = 10$ words for each topic and delete the duplicates to capture the topic words for each target, which achieves superior performance in the pilot studies.[3]. We apply early stopping in training process with $patience = 5$. The reported results are averaged scores of 10 runs to obtain statistically stable results[4].

4.2.2 Evaluation Metric. For Vast dataset, following [1], we perform Macro-averaged F1 of each label to measure the testing performance of the models. For Sem16 dataset, following [2], we report F_{avg}: the average of F1 on *Favor* and *Against*. For Wt-wt dataset, following [8], we report the Macro F1 score of each target.

4.3 Comparison Models

We compare our model with several strong baselines, including neural network-based method: `BiCond` [3], attention-based model: `CrossNet` [43], knowledge-based method: `SEKT` [45], graph network method: `TPDG` [23], adversarial learning method: `TOAD` [2], and BERT-based methods: `BERT` [9], `TGA Net` [1], `BERT-GCN` [24],

[2] Note that τ_p can be set to $\tau_p \in [0.05, 0.1]$, τ_y can be set to $\tau_y \in [0.1, 0.2]$, and α can be set to $\alpha \in [0.1, 0.5]$, which can also achieve state-of-the-art performance.

[3] In preliminary experiments, we set $T \in [1, 50]$ and $K \in [10, 50]$ to generate topic words, and found that the fluctuation of performance is negligible. Thus, the impact on performance is slight as long as the values are taken within a reasonable range.

[4] The source code of this work is released at https://github.com/HITSZ-HLT/PT-HCL.

Table 4: Experimental results on three ZSSD datasets. The results with $\natural$ are retrieved from [1], with $\dagger$ are retrieved from [24], with $\ddagger$ are retrieved from [2], with $\sharp$ are retrieved from [8], with $\flat$ are retrieved from [23], with $*$ indicates significance tests of our PT-HCL over BERT, TGA Net, CKE-Net, and PET (with $p < 0.05$). Best scores are in bold.

Model	VAST (%)				SEM16 (%)						WT-WT (%)			
	Pro	Con	Neu	All	DT	HC	FM	LA	A	CC	CA	CE	AC	AH
BiCond [3]	$44.6^{\natural}$	$47.4^{\natural}$	$34.9^{\dagger}$	$42.8^{\natural}$	$30.5^{\ddagger}$	$32.7^{\ddagger}$	$40.6^{\ddagger}$	$34.4^{\ddagger}$	$31.0^{\ddagger}$	$15.0^{\ddagger}$	$56.5^{\sharp}$	$52.5^{\sharp}$	$64.9^{\sharp}$	$63.0^{\sharp}$
CrossNet [43]	$46.2^{\natural}$	$43.4^{\natural}$	$40.4^{\dagger}$	$43.4^{\natural}$	35.6	38.3	41.7	38.5	39.7	22.8	$59.1^{\sharp}$	$54.5^{\sharp}$	$65.1^{\sharp}$	$62.3^{\sharp}$
SEKT [45]	$50.4^{\dagger}$	$44.2^{\dagger}$	$30.8^{\dagger}$	$41.8^{\dagger}$	-	-	-	-	-	-	-	-	-	-
TPDG [23]	53.7	49.6	52.3	51.9	47.3	50.9	53.6	46.5	48.7	32.3	$66.8^{\flat}$	$65.6^{\flat}$	$74.2^{\flat}$	$73.1^{\flat}$
TOAD [2]	42.6	36.7	43.8	41.0	$49.5^{\ddagger}$	$51.2^{\ddagger}$	$54.1^{\ddagger}$	$46.2^{\ddagger}$	$46.1^{\ddagger}$	$30.9^{\ddagger}$	55.3	57.7	58.6	61.7
BERT [9]	$54.6^{\natural}$	$58.4^{\natural}$	$85.3^{\dagger}$	$66.1^{\natural}$	$40.1^{\ddagger}$	$49.6^{\ddagger}$	$41.9^{\ddagger}$	$44.8^{\ddagger}$	$55.2^{\ddagger}$	$37.3^{\ddagger}$	$56.0^{\flat}$	$60.5^{\flat}$	$67.1^{\flat}$	$67.3^{\flat}$
TGA Net [1]	$55.4^{\natural}$	$58.5^{\natural}$	$85.8^{\dagger}$	$66.6^{\natural}$	40.7	49.3	46.6	45.2	52.7	36.6	65.7	63.5	69.9	68.7
BERT-GCN [24]	$58.3^{\dagger}$	$60.6^{\dagger}$	$86.9^{\dagger}$	$68.6^{\dagger}$	42.3	50.0	44.3	44.2	53.6	35.5	67.8	64.1	70.7	69.2
CKE-Net [24]	$61.2^{\dagger}$	$61.2^{\dagger}$	$88.0^{\dagger}$	$70.2^{\dagger}$	-	-	-	-	-	-	-	-	-	-
PET [32]	54.4	50.6	36.6	47.2	48.6	53.9	52.3	48.7	46.8	32.3	71.6	66.7	73.7	74.5
PT-HCL (ours)	**61.7***	**63.5***	**89.6***	**71.6***	**50.1***	**54.5***	**54.6***	**50.9***	**56.5***	**38.9***	**73.1***	**69.2***	**76.7***	**76.3***
-contrastive	57.6	59.7	87.3	68.2	43.4	50.8	44.5	45.5	54.9	37.6	69.3	66.1	73.9	72.5
w/ tf-idf	61.4	62.9	89.4	71.2	49.7	53.5	53.8	50.7	56.1	38.2	71.9	68.8	76.4	75.7
w/ similarity	61.1	63.2	89.0	71.1	49.3	54.2	54.4	49.8	55.7	38.6	72.6	67.7	75.9	76.0

and CKE-Net [24]. We also compare with a prompt-based method exploited in stance detection: Pattern-Exploiting Training (PET) [32].

In addition, we provide several variants of our proposed PT-HCL. "-contrastive" represents without using contrastive learning. This variant regards the pretext task and stance detection as a supervised multi-task learning. The Eq. 9 is replaced with: $\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_p + \lambda||\Theta||^2$. Here $\mathcal{L}_p$ is the cross-entropy loss of pretext task. "w/ tf-idf" represents using TF-IDF to generate target-related words paired with weights. "w/ similarity" denotes determining the target-related words and weights via the embedding similarity between words and the target. Based on it, we select the 10 words with the highest weight for each target as masked candidates.

Further, we also design various varieties of the proposed PT-HCL o analyze the impact of different components in the ablation study. "**w/o** *candidates*" denotes without deriving masked candidates to generate masked sentences. Inspired by [9], this variant randomly masks 15% of words in each sentence to generate masked sentences for the self-supervised learning pretext task. "**w/o** $\mathcal{L}_{type}$" denotes without using contrastive loss of pretext task to distinguish the types of stance features in the latent space. This variant only uses stance label as supervised signal in contrastive learning, i.e. Eq. 4 is replaced by Eq. 10. "**w/o** $\mathcal{L}_{stance}$" denotes without stance contrastive information in Eq. 4. "**w/o** $\mathcal{L}_{hierarchical}$" denotes without using hierarchical scenario, i.e. Eq. 4 is replaced by Eq. 11.

$$\ell(\boldsymbol{h}_i) = \log \frac{\sum_{j=1}^{N_b} \mathbb{1}_{[j \neq i]} \mathbb{1}_{[y^i = y^j]} f(\boldsymbol{h}_i, \boldsymbol{h}_j)}{\sum_{k=1}^{N_b} \mathbb{1}_{[k \neq i]} f(\boldsymbol{h}_i, \boldsymbol{h}_k)} \tag{10}$$

$$\begin{aligned} \ell(\boldsymbol{h}_i) = \log & \frac{\sum_{j=1}^{N_b} \mathbb{1}_{[j \neq i]} \mathbb{1}_{[p^i = p^j]} f(\boldsymbol{h}_i, \boldsymbol{h}_j)}{\sum_{k=1}^{N_b} \mathbb{1}_{[k \neq i]} f(\boldsymbol{h}_i, \boldsymbol{h}_k)} \\ + \log & \frac{\sum_{j=1}^{N_b} \mathbb{1}_{[j \neq i]} \mathbb{1}_{[y^i = y^j]} f(\boldsymbol{h}_i, \boldsymbol{h}_j)}{\sum_{k=1}^{N_b} \mathbb{1}_{[k \neq i]} f(\boldsymbol{h}_i, \boldsymbol{h}_k)} \end{aligned} \tag{11}$$

5 EXPERIMENTAL RESULTS

5.1 Main Experimental Results

We report the main experimental results of zero-shot stance detection on three benchmark datasets in Table 4. We observe that our PT-HCL performs consistently better than all baseline models on all datasets, which verifies the effectiveness of our proposed approach in ZSSD. Further, the significance tests of our PT-HCL over BERT, TGA Net, CKE-Net, and PET show that our PT-HCL presents a statistically significant improvement in terms of all evaluation metrics (with $p < 0.05$).

Furthermore, compared with the adversarial learning-based model (TOAD) that performs poorly in VAST and WT-WT datasets, our PT-HCL achieves outstanding performance. This demonstrates that method that exploits the discriminator in an adversarial learning way can not well identify the target information when the distribution of targets is imbalanced, while our PT-HCL can make good use of the differences and similarities between different targets in the light of the contrastive learning and thus leads to improved performance. In addition, in comparison with the vanilla BERT model, both our PT-HCL and *-contrastive* (the variant of our PT-HCL that without contrastive learning) achieve outstanding improvement. This indicates that exploring a self-supervised learning pretext task to derive supervised signal of target-invariant stance features is effective in leveraging shared stance features for the previously unseen targets, and thus leads to improved ZSSD performance. Further, our PT-HCL significantly outperforms *-contrastive* in terms of all evaluation metrics. This demonstrates the significance and validity of our proposed contrastive learning approach in ZSSD. According to the results of using different methods to derive masked candidates, we conclude that different methods lead to a slight impact on performance, and topic model used in our PT-HCL performs slightly better. This verifies the generalizability and effectiveness of our proposed method about determining target-related words for deriving masked candidates.

Table 5: Experimental results of ablation study.

Model	Vast (%)				Sem16 (%)						Wt-wt (%)			
	Pro	Con	Neu	All	DT	HC	FM	LA	A	CC	CA	CE	AC	AH
PT-HCL	**61.7**	**63.5**	**89.6**	**71.6**	**50.1**	**54.5**	**54.6**	**50.9**	**56.5**	**38.9**	**73.1**	**69.2**	**76.7**	**76.3**
w/o *candidates*	56.2	60.4	86.5	67.7	43.6	49.7	44.3	46.3	53.7	37.1	67.5	64.7	71.8	71.2
w/o $\mathcal{L}_{type}$	58.6	61.0	87.3	69.0	45.0	50.3	48.1	47.3	54.1	37.4	68.2	67.0	72.3	72.4
w/o $\mathcal{L}_{stance}$	60.6	62.3	88.4	70.4	49.6	51.7	52.3	48.9	55.7	38.3	71.3	68.6	75.5	74.8
w/o $\mathcal{L}_{hierarchical}$	59.2	61.7	88.5	69.8	46.2	51.6	47.8	46.6	54.4	37.8	69.9	67.1	74.2	73.0

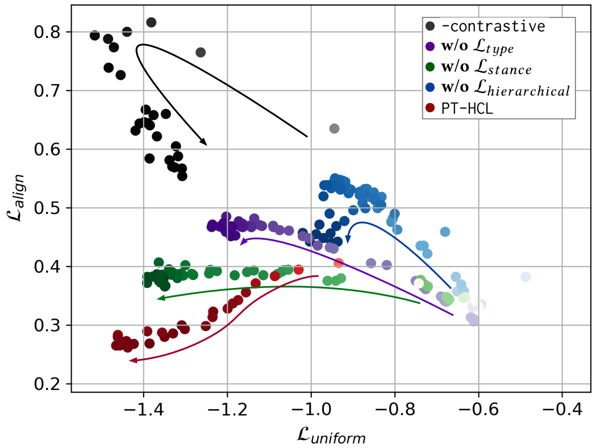


Figure 3: Visualization of contrastive representation of checkpoints from every 50 training steps. The darker the color of the point, the greater the accuracy. The arrows indicate the training direction. Models with low $\mathcal{L}_{align}$ and $\mathcal{L}_{uniform}$ consistently perform well (lower left corner).

5.2 Ablation Study

To investigate the impact of different components on performance, we conduct an ablation study of the proposed PT-HCL and report the results in Table 5. We observe that removal of masked candidates ("w/o *candidates*") sharply degrades the performance, which verifies the effectiveness and significance of deriving target-related words as masked candidates for each target, and thus capturing high-quality masked instances to perform a pretext task. From the results of "w/o $\mathcal{L}_{type}$", we conclude that distinguishing the types of stance features in the latent space are valid to improve the learning of transferable stance features in ZSSD. In addition, the performance declines considerably when the stance information is not considered in the contrastive loss ("w/o $\mathcal{L}_{stance}$"), which implies that the information of stance label is also an important supervised signal when performing contrastive learning. Further, note that removal of hierarchical scenario in contrastive loss ("w/o $\mathcal{L}_{hierarchical}$") leads to an evident decline in performance. This indicates that hierarchical contrastive loss is much superior in improving the quality of the learned feature representation. One possible reason is that placing pretext task and stance label on the same level in designing contrastive loss may lead to mutual conflict. As such, we explore a novel hierarchical contrastive loss to relieve the conflict between types and labels for better performance in ZSSD.

Table 6: Experimental results of few-shot scenario. Results of baselines are retrieved from [24].

Model	Pro	Con	Neu	All
BiCond [3]	45.4	46.3	25.9	39.2
Cross-Net [43]	50.8	50.5	41.0	47.4
SEKT [45]	51.0	47.9	21.5	47.4
BERT [9]	54.4	59.7	79.6	64.6
TGA Net [1]	58.9	59.5	80.5	66.3
BERT-GCN [24]	62.8	63.4	83.0	69.7
CKE-Net [24]	**64.4**	62.2	83.5	70.1
PT-HCL	62.3	**67.0**	**84.3**	**71.2**

5.3 Qualitative Analysis

5.3.1 *Analysis of Contrastive Representation.*

To further analyze how the proposed PT-HCL works in contrastive representation learning, we take the checkpoints from -contrastive, the three variants of PT-HCL, and our complete PT-HCL during training and visualize the alignment and uniformity metrics in Figure 3. Here, the evaluation and analysis of *alignment* and *uniformity* are following [40], which verifies that models attain both better alignment and uniformity will achieve better performance. From the results, we observe that our PT-HCL shows lower $\mathcal{L}_{align}$ and $\mathcal{L}_{uniform}$ during the training, which demonstrates that our PT-HCL attain strong ability in contrastive learning. Note that results of -contrastive present the worst alignment and uniformity, which indicates that contrastive learning can advance a better latent space for the learned representations. In addition, "w/o $\mathcal{L}_{hierarchical}$" leads to poorer metrics of both alignment and uniformity, which further verifies the effectiveness and significance of hierarchical strategy in our PT-HCL. Further, the results of "w/o $\mathcal{L}_{type}$" and "w/o $\mathcal{L}_{stance}$" are worse than the complete PT-HCL, which implies that both the types of stance features from the pretext and stance label are important in learning contrastive representation.

5.3.2 *Analysis of Few-Shot Scenario.*

Following [1, 24], we also evaluate our framework in the few-shot stance detection scenario on Vast dataset. The experimental results are shown in Table 6. Note that our PT-HCL performs overall better than all the comparison methods. This verifies the effectiveness and generalizability of our PT-HCL in dealing with both zero-shot and few-shot conditions.

5.3.3 *Analysis of Cross-Target Scenario.*

We further conduct comparison experiments in the cross-target scenario, a special form of zero-shot, on Sem16 dataset and report the results in Table 7. Note our PT-HCL framework achieves consistently better performance

Table 7: Experimental results of cross-target scenario. "HC→DT" denotes training on HC and testing on DT, etc. Results of baselines are retrieved from [23].

Model	HC→DT	DT→HC	FM→LA	LA→FM
BiCond [3]	29.7	35.8	45.0	41.6
CrossNet [43]	43.1	36.2	45.4	43.3
BERT [9]	43.6	36.5	47.9	33.9
SEKT [45]	47.7	42.0	53.6	51.3
TPDG [23]	50.4	52.9	58.3	54.1
PT-HCL	**53.7**	**55.3**	**59.3**	**54.6**

Table 8: Comparison results of combining our proposed PT-HCL framework with different stance detection models.

Model	Pro	Con	Neu	All
BiCond [3]	44.6	47.4	34.9	42.8
+PT-HCL	**52.1**	**57.5**	**45.8**	**51.8**
Cross-Net [43]	46.2	43.4	40.4	43.4
+PT-HCL	**56.5**	**60.1**	**52.4**	**56.3**
TGA Net [1]	55.4	58.5	85.8	66.6
+PT-HCL	**60.9**	**61.6**	**86.4**	**69.6**
BERT-GCN [24]	58.3	60.6	86.9	68.6
+PT-HCL	**63.6**	**64.8**	**87.8**	**72.1**

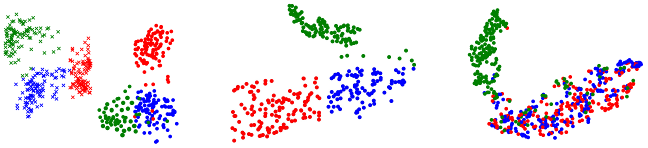
on all cross-target conditions, which verifies that our PT-HCL can generalize the superior learning ability to cross-target scenario. In addition, combining the results of Table 4, we observe that the results of cross-target scenario are overall better than ZSSD. This indicates that recognizing the relationships among targets in advance can potentially improve the stance detection performance for the unseen targets, and while showing the challenge of ZSSD.

5.3.4 Analysis of Different Encoders. In this section, we replace the BERT-based encoder in our PT-HCL with several strong neural network-based baselines and conduct experiments on Vast dataset. The comparison results are shown in Table 8. Note that all stance detection baselines can be directly combined with our proposed contrastive learning scenario and achieve outstanding improvement. This verifies the generalizability and effectiveness of the proposed contrastive learning strategy of our PT-HCL in ZSSD.

5.3.5 Analysis of the Pretext Task. We then make statistics on the prediction results of masked sentences on three datasets, and find that the probability intervals of correct results are 53%-56% on Vast dataset, 76%-85% on Sem16 dataset, and 67%-71% on Wt-wt dataset, respectively. This indicates that the prediction results of a considerable number of masked instances are unchanged, and thus those instances could be regarded as target-invariant stance expressions for learning transferable stance features in ZSSD.

5.4 Visualization

To qualitatively demonstrate how the proposed hierarchical contrastive learning strategy works in improving the quality of learned representations, we present the t-SNE [38] visualization of intermediate vectors learned by our PT-HCL and BERT on Vast dataset. The results are shown in Figure 4. Figure 4 (a) demonstrates that representations belonging to the same pretext label are pulled together, while the separation of representations between different pretext labels is quite clear. Further, within each embedding space of pretext label, the representation distributions among different stance labels are diverse. These results verify the effectiveness of the hierarchical contrastive learning in refining the distributions of learned representations. In addition, the visualizations of Figure 4 (b) and (c) present that the separations of testing data representations among different stance labels learned by our PT-HCL are considerably more apparent than the vanilla BERT. This implies that the PT-HCL can learn sounder inductive information from the training data owing to the hierarchical contrastive learning, so as to preferably generalize transferable stance features to the unseen targets and thus improve the performance of ZSSD.



(a) Training data (b) Testing data (c) BERT

Figure 4: Visualizations of intermediate vectors learned by our PT-HCL (a) and (b) and the vanilla BERT (c). Red=*Pro*, blue=*Con*, green=*Neu*. In (a), cross=*target-invariant*, dot=*target-specific*.

6 CONCLUSION

This paper presents a novel hierarchical contrastive learning framework based on the supervised information of labels and the surrogate supervised signal provided by a self-supervised learning pretext task in zero-shot stance detection (ZSSD), called PT-HCL. To be specific, we devise a novel hierarchical contrastive learning strategy to leverage the correlation and difference between target-invariant and target-specific stance representations, and further provide insights and discrimination into the distribution space of representations among different stance labels. This essentially allows the model to improve the quality of learned representations for learning transferable stance features in dealing with ZSSD. Experimental results on three benchmark datasets demonstrate that the proposed PT-HCL consistently outperforms the state-of-the-art baseline models in ZSSD task.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (61876053, 62006062, 62176076), UK Engineering and Physical Sciences Research Council (grant no. EP/V048597/1, EP/T017112/1), Natural Science Foundation of Guangdong Province of China (No. 2019A1515011705), Shenzhen Foundational Research Funding (JCYJ20180507183527919, JCYJ20200109113441941, JCYJ20210324115614039), Shenzhen Science and Technology Innovation Program (Grant No. KQTD20190929172835662), Joint Lab of HITSZ and China Merchants Securities. Yulan He is supported by a Turing AI Fellowship funded by the UK Research and Innovation (UKRI) (grant no. EP/V020579/1).

REFERENCES

[1] Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8913–8931.

[2] Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial Learning for Zero-Shot Stance Detection on Social Media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4756–4767.

[3] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 876–885.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. 1597–1607.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 22243–22255. https://proceedings.neurips.cc/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf

[7] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. 2019. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12154–12163.

[8] Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1715–1724.

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

[11] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance Classification with Target-specific Neural Attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3988–3994.

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*.

[13] 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. https://aclanthology.org/2021.acl-long.72

[14] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, suppl 1 (2004), 5228–5235.

[15] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=cu7IUiOhujH

[16] Kazi Saidul Hasan and Vincent Ng. 2014. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 751–762.

[17] Mohammad Kachuee, Hao Yuan, Young-Bum Kim, and Sungjin Lee. 2021. Self-Supervised Contrastive Learning for Efficient User Satisfaction Prediction in Conversational Agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4053–4064.

[18] Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. tWT–WT: A Dataset to Assert the Role of Target Entities for Detecting Stance of Tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3879–3889.

[19] Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4725–4735.

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 18661–18673.

[21] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. 2016. Learning representations for automatic colorization. In *European conference on computer vision*. Springer, 577–593.

[22] Yingjie Li and Cornelia Caragea. 2019. Multi-Task Stance Detection with Sentiment and Stance Lexicons. 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)*. 6299–6305.

[23] Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive Graph for Cross-target Stance Detection. In *the Web Conference 2021 (WWW '21)*.

[24] Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing Zero-shot and Few-shot Stance Detection with Commonsense Knowledge Graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3152–3157.

[25] Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 1065–1072.

[26] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6707–6717.

[27] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 31–41.

[28] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*. Springer, 69–84.

[29] 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.

[30] Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning. 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)*. 3350–3363.

[31] Yao Qiu, Jinchao Zhang, and Jie Zhou. 2021. Improving Gradient-based Adversarial Training for Text Classification by Contrastive Learning and Auto-Encoder. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1698–1707.

[32] Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 255–269.

[33] Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet Stance Detection Using an Attention based Neural Ensemble Model. 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)*. 1868–1873.

[34] Nathaniel Simard and Guillaume Lagrange. 2021. Improving Few-Shot Learning with Auxiliary Self-Supervised Pretext Tasks. *arXiv preprint arXiv:2101.09825* (2021).

[35] Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. 116–124.

[36] Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance Detection with Hierarchical Attention Network. In *Proceedings of the 27th International Conference on Computational Linguistics*. 2399–2409.

[37] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 776–794.

[38] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.

[39] 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)*. 2495–2504.

[40] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*. PMLR, 9929–9939.

[41] Penghui Wei and Wenji Mao. 2019. Modeling Transferable Topics for Cross-Target Stance Detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1173–1176.

[42] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3733–3742.

[43] Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-Target Stance Classification with Self-Attention Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 778–783.

[44] Nan Yang, Furu Wei, Binxing Jiao, Daxing Jiang, and Linjun Yang. 2021. xMoCo: Cross Momentum Contrastive Learning for Open-Domain Question Answering. 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)*. 6120–6129.

[45] Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3188–3197.

[46] Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021. Supporting Clustering with Contrastive Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5419–5430.

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