Generative Data Augmentation with Contrastive Learning for Zero-Shot Stance Detection

Yang Li
Northeast Forestry University,
Harbin 150004, China
yli@nefu.edu.cn

JiaWei Yuan
Northeast Forestry University,
Harbin 150004, China
jwyuan@nefu.edu.cn

Abstract

Stance detection aims to identify whether the author of an opinionated text is in favor of, against, or neutral towards a given target. Remarkable success has been achieved when sufficient labeled training data is available. However, it is labor-intensive to annotate sufficient data and train the model for every new target. Therefore, zero-shot stance detection, aiming at identifying stances of unseen targets with seen targets, has gradually attracted attention. Among them, one of the important challenges is to reduce the domain transfer between seen and unseen targets. To tackle this problem, we propose a generative data augmentation approach to generate training samples containing targets and stances for testing data, and map the real samples and generated synthetic samples into the same embedding space with contrastive learning, then perform the final classification based on the augmented data. We evaluate our proposed model on two benchmark datasets. Experimental results show that our approach achieves state of-the-art performance on most topics in the task of zero-shot stance detection.

Target-specific stance detection	
Training Step	**Target**: climate change is a real concern
Sentence: When your wearing sweaters in the summer.	**Stance**: Pro
Test Step	**Target**: climate change is a real concern
Sentence: Netherlands just taught the rest of the world a very important lesson.	**Stance**: Pro
Zero-shot stance detection	
Training Step	**Target**: climate change is a real concern
Sentence: When your wearing sweaters in the summer.	**Stance**: Pro
Test Step	**Target**: atheism
Sentence: Take destiny hands people place hands god.	**Stance**: Con

Figure 1: An Illustration of Target Specific Stance Detection and Zero-Shot Stance Detection. In ZSSD, the annotation data of the target to be tested will not appear in the training step of the model.

1 Introduction

In recent years, social media websites have become an important platform for people to express their opinions on different targets (or topics in some literature) ranging from politics, government policies, movies, sports and social issues, etc (ALDayel and Magdy, 2021). More often, users tend to take a stance, in Favor, Against or Neutral towards a particular target (Mohammad et al., 2016). The task of stance detection aims to automatically identify the stance represented by the users in numerous texts. It is of great significance to applications like argument mining, fake news detection, public opinion analysis and has caught the attention of researchers.

Remarkable success has been achieved when sufficient labeled training data is available in the task of stance detection (Sun et al., 2018; Siddiqua et al., 2019a; Du et al., 2020). Most of the existing research work relies on a large number of labeled data for a specific target. However, in reality, new targets are constantly produced, and it is unpractical to label sufficient data for each target. Based on the idea of transfer learning, some researchers use cross-target stance detection to train one classifier adapting from the current target to "new" target (Sobhani et al., 2017). However, this adaptation depends on the correlation between targets, and still needs annotation data of "new" targets. To solve the problem of missing annotating data of "new" targets in the real world scenario, the task of Zero-Shot Stance Detection (ZSSD), aiming at classifying stances for a large number of unseen targets without training data, has emerged (Allaway and Mckeown, 2020).

A great challenge in the task of ZSSD is the generalization ability of the model due to the lack of labeled data of specific targets. Some studies try to introduce external knowledge (*e.g.* Knowledge Graph) to capture the correlation between targets (Du et al., 2017; Liu et al., 2021). However, due to the existence of domain specific features, directly transferring stance features from seen targets to unseen targets may not result in good performance.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6985–6995
December 7-11, 2022 ©2022 Association for Computational Linguistics

To tackle this problem, we propose generating high-quality training data for unseen targets based on the training data of seen targets. In order to ensure the quality of the generated samples, we use the discriminator and generator in Generative Adversarial Networks (GANs) for adversarial learning (Goodfellow et al., 2014; Meng et al., 2022). In addition, we conduct hybrid contrastive learning on the synthesized samples and the ground-truth target samples(Allaway and Mckeown, 2020), to remove the noise data irrelevant to the target and improve the quality of generated samples (Siddiqua et al., 2019b). Through the instance-level and the class-level contrastive learning, the knowledge in different spaces can be effectively transferred. Experimental results on two public ZSSD data sets show that our method can significantly outperform various competitive baselines. The ablation experiment proves the effectiveness of each part of our proposed model.

The main contributions of our work can be summarized as follows:

- We propose a generative data augmentation model which generates high-quality training data for unseen targets by adversarial learning and contrastive learning to improve the performance of zero-shot stance detection.
- We design comprehensive experiments on two ZSSD benchmark datasets to compare our approach with the-state-of-art baselines. Experimental results show that our approach outperforms the baselines.
- To the best of our knowledge, this is the first time to use the generative data augmentation method to improve the task of ZSSD, without any prior knowledge.

2 Relate Work

Existing research on stance detection can be roughly divided into target-specific and cross-target stance detection (Augenstein et al., 2016; Du et al., 2017). For target-specific stance detection, most studies rely on a large number of labeled data for the specific target to train classifiers through different deep learning models (Siddiqua et al., 2019b; Darwish et al., 2020; Sun and Li, 2021; Kawintiranon and Singh, 2021). While cross-target stance detection is based on the correlations between user's stances on different targets. Researchers proposed many approaches, such as attention-based model (Xu et al., 2018; Wei and Mao, 2019), memory-based model (Wei et al., 2018) and graph-based model (Liang et al., 2021), to capture the underlying knowledge transferred between targets. Further, (Li and Caragea, 2021) proposed a novel data augmentation approach by predicting the masked token and replacing a mentioned target with another that achieved promising performance on Multi-Target stance detection.

Unlike the above tasks, zero-shot stance detection aims learning a classifier that is evaluated on a large number of completely new targets. Allaway and Mckeown (2020) introduced a new dataset of news article comments for zero-shot stance detection and proposed a Topic-Grouped Attention model to implicitly construct relationships between the seen and unseen targets. Inspired by adversarial learning for domain adaptation, Allaway et al. (2021) extracted target-invariant transformation features by adversarial learning. In addition, by introducing commonsense knowledge, Liu et al. (2021) make use of the structural-level and semantic-level information of relational subgraphs to strengthen generalization capability of the model.

The above work focused on using existing features to embed a text and any possible attribute description into their corresponding latent representations. The main goal of this embedding based approach is to map textual features and attribute descriptions into a common embedding space by using projection functions, which are learned by deep networks (Fu et al., 2017; Kawintiranon and Singh, 2021). However, the method based on embedding is more inclined to predict the seen class labels as the output, it will cause the problems of distribution deviation and domain shift. In order to overcome this problem, some works generate training data for the unseen classes through the generation model(Verma et al., 2018; Huang et al., 2019; Han et al., 2021; Schick and Schütze, 2021). By augmenting the data of the target domain, zero-shot classifier can be trained on all samples of known and unknown classes. Our work is based on these considerations.

3 Methodology

In this section, before introducing our proposed Generative Data Augmentation framework with Contrastive Learning (GDA-CL for short), we will first define the task of ZSSD. The overall model is shown in Figure 2, which consists of three parts, namely Training Data Generation, Hybrid Con-

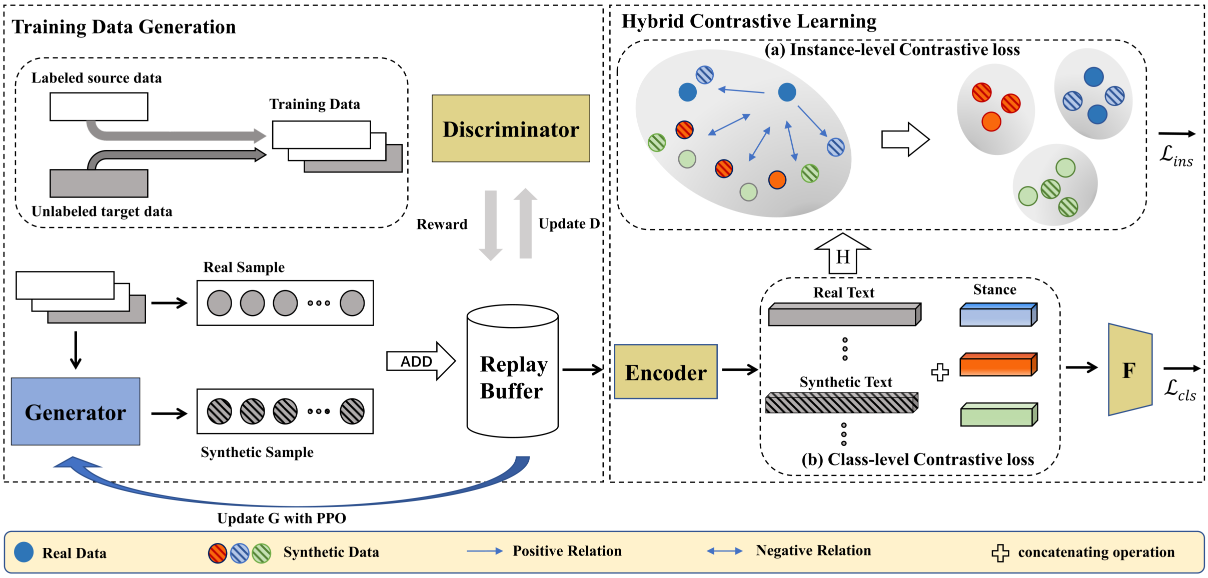


Figure 2: The architecture of the proposed GDA-CL framework.

trastive Learning and Fine-Tuning in Stance Classification.

3.1 Problem Formulation

Formally, in the task of zero-shot stance detection, we have two disjoint sets: S for seen targets t_s and U for unseen targets t_u, where $t_s \cap t_u = \varnothing$. Suppose that there are N_s labeled instances $\mathcal{D}_s = \left\{\left(x_s^i, t_s^i, y_s^i\right)\right\}_{i=1}^{N_s}$ provided for training, where $x_s^i = \left[x_i^1, x_i^2, \ldots, x_i^l\right]$ is a sequence of l words, representing the user's stance text in this task, t_i is the corresponding target and y_i is the stance label. The test set $\mathcal{D}_u = \left\{\left(x_u^i, t_u^i, y_u^i\right)\right\}_{i=1}^{N_u}$ contains N_u unlabeled instances. The objective of ZSSD task is to predict a stance label for each x_i towards unseen target t_u^i in $\mathcal{D}_u$, based on the model trained from each x_i towards the seen target t_s^i in $\mathcal{D}_s$.

3.2 Training Data Generation

To learn transferable stance features from the seen targets, we generate synthetic targeted training data $\mathcal{D}_g = \left\{\left(\hat{x}_1, \hat{y}_1, \hat{y}_1\right), \cdots, \left(\hat{x}_n, \hat{y}_n, \hat{t}_n\right)\right\}$ for unseen targets from the original targeted training data $\mathcal{D}_s$ using Pretraining Language Model (PLM) and contrastive learning to assist text generation.

We introduce a text generation model based on generative adversarial imitation learning (GAIL) (Ho and Ermon, 2016) to synthesize the missing training samples for unseen targets. The framework consists of a generator G_θ and a discriminator D_ϕ, which are parameterized with θ and ϕ, respectively. For each given training instance $d = (x, t, y)$, we concatenate t and y with prefixes as text generation prompt a like "The topic is focus on [t], the label is [y]" to control targets and labels of the synthetic samples. The combination of t and y acts as the semantic description of the target in our framework. We use GPT-2 (Radford et al., 2019) as a generator network G_θ to produce the samples $\hat{x} = G(a)$ conditioned on an attribution description a. At the same time, Roberta (Liu et al., 2019) is used as discriminator D_ϕ to distinguish a ground-truth sample x from a synthetic sample $\hat{x}$. As a result, each synthesized sample $\hat{x}$ will obtain a single sparse reward. The saddle point of model is where the generator and discriminator satisfy the following objective function at the same time:

$$\min_{G_\theta} \max_{D_\phi} \mathbb{E}_{p_{real}} \left[D_\phi(a, x)\right] + \mathbb{E}_{G_\theta} \left[1 - D_\phi\left(a, G_\theta(a)\right)\right] \quad (1)$$

The discriminator outputs confidence scores p_r and p_g between 0 and 1 for ground-truth sequences and the generated sequences, respectively. Then the confidence scores are used to optimize the discriminator through cross-entropy loss and provide a reward signal R_y for the generator. In the training process of the generator, the action space is the whole vocabulary, so we use imitation replay algorithm to ensure stability inspired by the recent works (Gulcehre et al., 2019) and (Reddy et al., 2019). Here, we set a ratio λ (*e.g.*, 0.3) for replay buffer to control the proportion of ground-truth sequences and generated sequences. In the hybrid buffer, generated sequences obtain reward from the discriminator D_ϕ, the scores of ground-truth sequences are set to a constant. However, the scores estimate will be very noisy due to the discriminator not being fixed, so they can not always predict the exact value of that state. Because the discriminator has been continuously optimized, the predicted scores fluctuate, which may cause the generator model to change too much. To tackle this problem, we choose to use simple and effective proximal policy optimization (PPO) (Schulman et al.) rather than trust region policy optimization (TRPO) (Schulman et al., 2015) to ensure $G_{\theta_{i+1}}$ not move far away from G_θ.

With PPO strategy, we calculate the current and last strategy likelihood ratio of the generator model G_θ:

$$r(\theta) = \frac{G_\theta\left(y_{1:T} \mid a\right)}{G_{\theta_{old}}\left(y_{1:T} \mid a\right)} \tag{2}$$

$$L_{\mathrm{G}}(\theta) = -\min\begin{cases} r(\theta)\hat{R}_y \\ clip(r(\theta), 1-\epsilon, 1+\epsilon)\hat{R}_y \end{cases} \tag{3}$$

where $y_{1:T}$ is a text sequence, T is the sequence length, ϵ is a clip factor, $\hat{R}_y$ is the reward after R_y normalization.

3.3 Hybrid Contrastive Learning

We introduce two levels of contrastive learning to guide the optimization of the generator. For the instance-level, we utilize contrastive loss to force the generated text closer to the ground-truth text with the same stance label. While the class-level contrastive loss gives higher scores to the samples that are consistent with the ground-truth labels.

The embedding of each sample d_i is encoded by an encoder to obtain the corresponding representation h_i, and then we feed it into the projection function: $z_i = H\left(h_i\right) = H\left(E\left(x_i\right)\right)$. In instance-level contrastive learning, we divide the samples into positive and negative samples according to their different stance labels. For each embedding z_i, The positive sample z^+ having the same stance label with z_i is selected, while the stance labels of the negative samples are different from z_i. Concretely, the cross-entropy loss is calculated as follows:

$$s_{ins}\left(z_i, z\right) = \exp\left(z_i * z / \tau_e\right) \tag{4}$$

$$\ell_{ins}\left(z_i, z^+\right) = -\log\frac{s_{ins}\left(z_i, z^+\right)}{\sum_{k=1}^{K} s_{ins}\left(z_i, z_k\right)} \tag{5}$$

where $\tau_e > 0$ is the temperature parameter for the instance-level contrastive embedding.

Similarly, we divide the data into positive and negative samples according to the similarities and differences of their textual descriptions in class-level contrastive embedding. We introduce a contrastive network $F(h, a)$ that computes the correlation score between an embedding h and a semantic description a. The class-level contrastive loss is defined as follows:

$$s_{cls}\left(h_i, a\right) = \exp\left(F\left(h_i, a\right) / \tau_s\right) \tag{6}$$

$$\ell_{cls}\left(h_i, a^+\right) = -\log\frac{s_{cls}\left(h_i, a^+\right)}{\sum_{s=1}^{S} s_{cls}\left(h_i, a_s\right)} \tag{7}$$

where $\tau_s > 0$ is the temperature parameter.

Finally, we integrate the instance-level contrastive loss L_{ins} and class-level contrastive loss L_{cls} into GANs. To train the discriminator, we can directly combine the contrastive loss with the classification loss.

$$L_D' = L_D + L_{ins}(x, t) + L_{cls}(x, t, y) \tag{8}$$

We combine the contrastive loss with the reward in the training process of generator. Hence, the reward in Equation 3 is updated as follows:

$$\hat{R}_y' = \hat{R}_y - L_{ins}(x, t) - L_{cls}(x, t, y) \tag{9}$$

3.4 Fine-Tuning in Stance Classification

As the generator model is trained on the labeled data of seen targets, the generated texts x^g may contain some noise unrelated to the unseen target. Therefore, directly applying all the generated data D_g to the training classifier C may lead to the prediction biases. In this section, we make a data selection strategy to the generated data, and train the classifier with the filtered data set.

3.4.1 Data Selection

The aim of data selection is to keep more information related to target t and label y in the generated text x^g. We regard the generated probability, which is produced by G_θ conditioned on the attribution description a, as the confidence score of the generated text. To eliminate the influence of different text lengths, we use the average log probability of all tokens x^g of as the score function s, which is similar to previous study (Yuan et al., 2021).

$$s = \frac{1}{n}\sum_{i=1}^{n}\log p_\theta\left(x_i \mid \left[\boldsymbol{a}; \boldsymbol{x}_{<i}^g\right]\right) \tag{10}$$

Given $K * N$ generated texts, we sort them according to the scores and keep the top N high probability samples as $\hat{D}_g$.

Statistics	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 1: Statistics of VAST dataset.

Target	Pro	Con	Neu	Keywords
DT	148	299	260	trump
HC	163	565	256	hillary,clinton
FM	268	511	170	femini
LA	167	544	222	aborti
CC	335	26	203	climate
A	124	464	145	atheism,atheist

Table 2: Statistics of SemT6 dataset.

3.4.2 Objective Function

We use the Pre-trained Language Models T_{LM} as classification model, then fine-tune the model by minimizing the cross-entropy loss with label smooth (Szegedy et al., 2016) to optimize the model. Note that our training data D_{train} is composed of two parts: the selected generated samples for unseen target $\hat{D}_g$ and the original training samples D_s.

4 Experiments

In this section, we perform experiments to answer the following research questions: RQ1. Can the generative data augmentation approach effectively improve the performance of zero-shot stance detection? If so, how much improvement is our method compared with other baselines? RQ2. Can Generative Adversarial Networks (GANs) and contrastive learning help to improve the quality of the generated texts? RQ3. How sensitive is our method to the parameters?

4.1 Datasets and Evaluation Metrics

The dataset used in this paper consists of two benchmark datasets.

VAST (Allaway and Mckeown, 2020): This dataset includes a large amount of specific targets (topics). The targets in VAST are diverse and the training data of each target is very small, which makes it very suitable for zero-shot stance detection task. The statistics of VAST dataset are shown in Table 1.

SemT6 (Allaway et al., 2021): This dataset contains six targets obtained from English social media. Specifically, it includes Donald Trump (DT), Hillary Clinton (HC), Feminist Movement (FM), Legalization of Abortion (LA), Climate Change (CC), and Atheism (A). Each instance has a stance label as Pro, Con or Neu. For the task of zero-shot stance detection, we select five targets as training dataset and the remaining one as test dataset. The statistics of SemT6 dataset are shown in Table 2.

Following the previous work (Allaway and Mckeown, 2020), for the VAST dataset, we use the macro average of F1-score as the evaluation metric, and for SemT6 dataset, we use the average F1-score F_{avg} of the class `Pro` and `Con` (Allaway et al., 2021).

4.2 Experimental Settings

We employ the basic version of BERT (Kenton and Toutanova, 2019) with 768-dimensional embedding as the classifier C. For the encoder E, we use the Roberta-base model (Liu et al., 2019). We use TextGAIL (Wu et al., 2021) as the generator G and discriminator D. The contrastive network F is a multi-layer perceptron (MLP) containing one hidden layer. In the generation process of training data, we use the AdamW optimizer (Loshchilov and Hutter, 2018) with the learning rate $lr = 1e^{-5}$ and L_2-regularization $\lambda = 1e^{-5}$. The proportion of data selection K is set to 10%. For SemT6, the classifier is optimized by the Adam optimizer(Kingma and Ba, 2015) with a learning rate of $1e^{-3}$. For VAST, we fine tune the whole BERT model with a learning rate of $2e^{-5}$.

4.3 Comparison Models

We compare our model with the several state-of-the-art baselines, including **BiCond** (Augenstein et al., 2016): bidirectional conditional encoding model, **CrossNet** (Xu et al., 2018): BiCond with topic-specific self-attention, **SKET** (Zhang et al., 2020): using a knowledge graph to transfer target features, **TOAD** (Allaway et al., 2021): domain adaptation of different targets through adversarial learning. We also compare with five Bert-base models **Bert** (Kenton and Toutanova, 2019), **TGA Net** (Allaway and Mckeown, 2020): using contextual conditional encoding and topic-grouped attention, **Bert-GCN**(Liu et al., 2021): BERT based on Graph Convolution Networks (GCN), and **CKE-Net** (Liu et al., 2021): commonsense knowledge enhanced Bert based on GCN.

Model	VAST				SemT6					
	Pro	Con	Neu	All	DT	HC	FM	LA	A	CC
Bicond	.459	.475	.349	.427	.305	.327	.406	.344	.310	.150
CrossNet	.462	.434	.404	.434	.356	.383	.417	.385	.397	.228
SEKT	.504	.442	.308	.418	-	-	-	-	-	-
TOAD	.426	.367	.438	.410	.495	.512	**.541**	.462	.461	.309
BERT	.546	.584	.853	.661	.401	.496	.419	.448	**.552**	.373
TGA Net	.554	.585	.858	.666	.407	.493	.466	.452	.527	.366
BERT-GCN	.583	.606	.869	.686	.423	.500	.443	.442	.536	.355
CKE-Net	**.612**	.612	.880	.702	-	-	-	-	-	-
GDA-CL	.598	**.623**	**.893**	**.705**	**.503**	**.554**	.534	**.475**	**.438**	.437

Table 3: Experimental results on two ZSSD datasets.

Model	VAST				SemT6					
	Pro	Con	Neu	All	DT	HC	FM	LA	CC	A
GDA-CL	.598	**.623**	**.893**	**.705**	**.503**	**.554**	**.534**	**.475**	**.437**	**.438**
(w/o CLS)	.602	.607	.882	.697	.501	.542	.531	.472	.422	.433
(w/o INS)	**.612**	.598	.884	.698	.499	.538	.526	.468	.422	.431
(w/o label smooth)	.583	.605	.877	.688	.447	.528	.506	. 455	.434	.406
(w/o data selection)	.611	.570	.890	.690	.487	.508	.490	.428	.378	.371

Table 4: Ablation experimental results on two ZSSD datasets.

5 Results

5.1 Main Experimental Results

In table 3, we compare our model GDA-CL with the competitive baseline methods on two benchmark datasets. For the VAST data, it has 4033 zero-shot topics, while SemT6 has six targets and provides high-quality texts for each topic. These data are able to support the training of the model. The following conclusions can be drawn from the experiment results: (i) Our model GDA-CL outperforms TOAD, BERT, BERT-GCN and TGA-NET, and achieves state-of-the-art results in VAST dataset and four targets (DT, HC, LA, A) in SemT6 dataset. (ii) When our model compare with Bicond-based model, we observe CrossNet and TOAD both obtain acceptable results on SemT6, but perform poor on VAST. This is because it is not applicable in the dataset composed of a large number of different targets by distinguishing the correlation of targets and constructing the relationship graph between targets. However, our method does not depend on the correlations between targets, and directly generates training data of unseen targets, which can be adapted to complex real-world scenarios. (iii) Additionally, we observe that our model GDA-CL has achieved satisfactory advantage and outperformed Bert-GCN and CKE networks, both of which use Bert-based graph neural networks. The improvement of our model comes from data augmentation, and no additional external knowledge is introduced, which verifies the effectiveness of our model.

5.2 Ablation Study

We conduct ablation study for different components of our model including class-level contrastive embedding, instance-level contrastive embedding, smooth loss and data selection on the VAST and SemT6 benchmark under the setting of zero-shot stance detection. In Table 4, the model *w/o CLS* and *w/o INS* represent our model GDA-CL without class-level contrastive embedding and instance-level contrastive embedding, respectively. It is found that after contrastive learning is removed, the results on both data sets have declined. For *w/o label smooth*, we directly use cross entropy loss without label smooth strategy in stance classification. The experimental results show that the label smoothing strategy to prevent over-fitting, which makes the model have higher generalization ability. Furthermore, *w/o data selection* represents the model without data selection. We found that the experimental results decreased significantly either without label smoothing or data selection. It demonstrates that the quality of the generated samples is very important to our method.

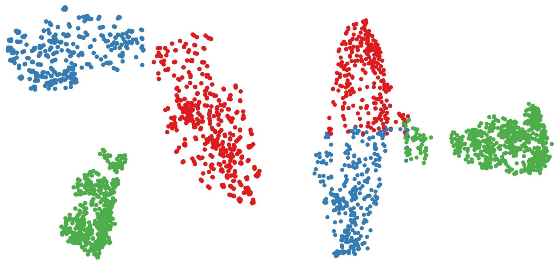
SemT6	Example 1	Example 2
Target	Climate change is a real concern	Legalization of abortion
Ground Truth	When your wearing sweaters in the summer.	Remind love means willing give hurts mother.
Generated Texts	(1) Couldn't believe place would keep people safe if climate change. (2) God help us calm climate arctic.	(1) Two years ago abortion was illegal. (2) Let legalize make America great.

Table 5: Examples generated by our model GDA-CL in SemT6 dataset. "Ground Truth" represents the ground-truth training samples of unseen targets, and "Generated Texts" represents the generated training samples of unseen targets.

VAST	Example 1	Example 2
Target	Constitutional amendment	Gun death
Ground Truth	Constitutional amendment make voting right. Currently disenfranchise citizens residing abroad vote state elections colleges ...	Professor sidesteps reality going nation 30,000 gun deaths per year many many multiples times gun related deaths per capita developed nations ...
Generated Texts	Realists learned much about outlaw competition, america bound cities govern big business. Nothing makes smaller firms compete cheap easier in mass market society.	So madness. learned clearly support legislation remove ban gun lead multiple victims deaths.

Table 6: Examples generated by our model GDA-CL in VAST dataset.

(a) Visualization with and without class-level contrastive learning



(b) Visualization of instance-level contrastive learning (c) Visualization of class-level contrastive learning

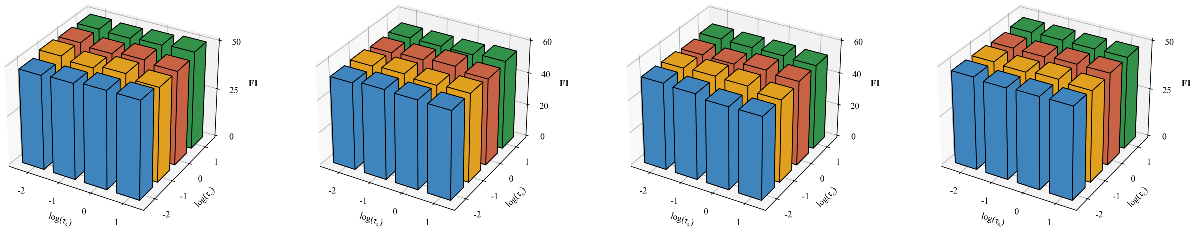
Figure 3: Visualization of intermediate embeddings from instance-level and class-level contrastive learning. Dots in different colors indicate samples belonging to different stance labels. Blue=Pro, Red=Con, Green=Neu.

5.3 Data Visualization

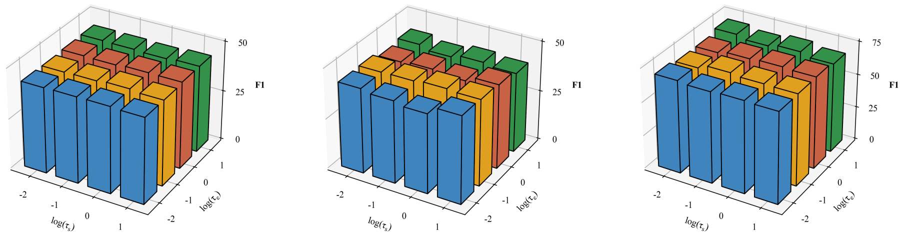
To further analyze the effect of contrastive learning, we visualize the intermediate representation vectors of text samples produced by the model through the visualization tool t-SNE(Van der Maaten and Hinton, 2008). Figure 3(a) shows the comparison of visualized embeddings learned with contrastive learning component and without contrastive learning component in training data. We can observe that class-level contrastive learning obviously pulls the representations belonging to the same label (same color) together, the representations between different labels are pulled away. From the visualization of instance-level and the class-level contrastive embeddings in Figure 3(b) and Figure 3(c), we can find that the distribution of representations belonging to different stance is separate in test dataset. This is consistent with our experimental results, showing that contrastive learning can guide the optimization of the data generation in GANs.

5.4 Qualitative Analysis

In Table 5 and 6, we show some training examples generated by GDA-CL for unseen targets in the SemT6 dataset and VAST dataset, respectively.



(a) F1-score for DT (b) F1-score for HC (c) F1-score for FM (d) F1-score for LA



(e) F1-score for CC (f) F1-score for A (g) F1-score for VAST

Figure 4: The results of F1-score in ZSSD with respect to different temperature parameters τ_e and τ_s. With the different τ_e and τ_s values, the F1-score results on different datasets change slightly, indicating that our method is robust to the temperature parameters.

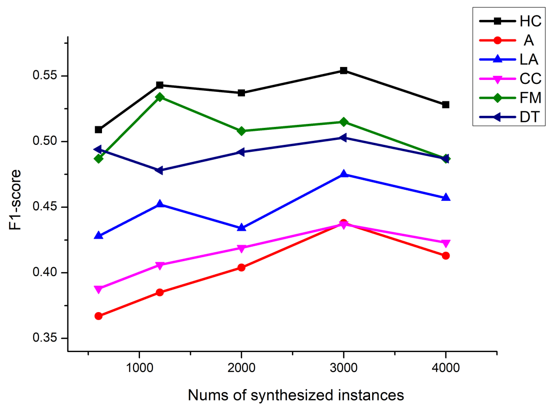


Figure 5: The results of F1-score in ZSSD with respect to different nums of synthesized instances.

From Table 6, we can observe that the average length of the training texts are usually less than ten words so they are easier to be learned. For each target, we show two generated sample texts with different stance labels. The generated training texts have similar semantics to the ground-truth samples, and are fluent. For VAST, we can observe that as the pre-training language model GPT2 has strong generalization ability on large-scale data, there are some new words like "Realists" and "legislation" generated in the samples, which have not appeared in training dataset.

5.5 Parameter Sensitivity Analysis

Next, we conduct parameter sensitivity analysis of our model. We first evaluate the F1-scores of our model under different temperature parameters in contrastive learning. In Figure 4, we set the two parameters τ_e and τ_s to $[0.01, 0.1, 1, 10]$ separately and show the performances on different targets. For most targets in SemT6 dataset, the best results are obtained when $\tau_e = 0.1$ and $\tau_s = 1$. In the VAST dataset, the best result is achieved when $\tau_e = 0.1$ and $\tau_s = 0.1$. All the experimental results show that our model is relatively stable when the parameters are changed.

Next, we try to add different numbers of generated samples to the training set in SemT6 dataset. The experimental results in Figure 5 show that when $N = 3000$, our model achieves the best performance in the task of zero-shot stance detection towards most targets.

6 Conclusion

In this article, we propose a generative data augmentation model that generates high-quality training data for unseen targets by adversarial learning and contrastive learning, for the task of zero-shot stance detection. We conducted a series of experiments to evaluate our approach against several state-of-the-art models on two benchmark datasets and found that our method outperforms the baselines significantly. Through qualitative analysis and visual analysis, we show that the generated texts for unseen targets have good fluency while maintaining semantics.

Limitations

In our work, we conducted experiments on two datasets. Compared to the VAST dataset, we achieve a greater improvement in SemT6 dataset. We suppose that one possible reason is that the overall average length of samples on VAST is larger than that in SemT6.

Although our existing model has a good performance in generating coherent short texts. For long text like paragraphs, it is difficult to dynamically model the input data, and it is also difficult to perfectly capture the complex semantics of long texts. This leads to an inherent limitation of our model in dealing with long texts.

Acknowledgments

We thank the anonymous reviewers for their insightful comments, and gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via Grant 62276059. We thank Qingfu Zhu for his valuable comments. Corresponding author: Yang Li, E-mail: yli@nefu.edu.cn.

Ethics Statement

In this work, two datasets used for task evaluation were obtained in the following ways. The datasets SemT6 and VAST are both downloaded directly. The SemT6 dataset is developed by a series of international natural language processing (NLP) research seminars in shared tasks. They allow the use of copyrighted material for research purposes without the permission of the copyright owner. The VAST dataset has an explicit statement from its author "We make our dataset and models available for use".

Some methods we discussed in this article include predicting the stance labels of some sensitive topics(*e.g.* politics, feminism). The use of these models may lead to unreasonable results due to misinformation, so these predicted results cannot be regarded as people's opinions in real life.

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

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)*, pages 8913–8931.

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*, pages 4756–4767.

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*, pages 876–885.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

Jiachen Du, Lin Gui, Ruifeng Xu, Yunqing Xia, and Xuan Wang. 2020. Commonsense knowledge enhanced memory network for stance classification. *IEEE Intelligent Systems*, 35(4):102–109.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. 2017. Zero-shot learning on semantic class prototype graph. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):2009–2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Caglar Gulcehre, Tom Le Paine, Bobak Shahriari, Misha Denil, Matt Hoffman, Hubert Soyer, Richard Tanburn, Steven Kapturowski, Neil Rabinowitz, Duncan Williams, et al. 2019. Making efficient use of demonstrations to solve hard exploration problems. In *International conference on learning representations*.

Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2371–2381.

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

He Huang, Changhu Wang, Philip S Yu, and Chang-Dong Wang. 2019. Generative dual adversarial network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 801–810.

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*, pages 4725–4735.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860.

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 *Proceedings of the Web Conference 2021*, pages 3453–3464.

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*, pages 3152–3157.

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.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding.

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)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Siddharth Reddy, Anca D Dragan, and Sergey Levine. 2019. Sqil: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. Association for Computational Linguistics.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019a. 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)*, pages 1868–1873, Minneapolis, Minnesota. Association for Computational Linguistics.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019b. 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)*, pages 1868–1873.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

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*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuqing Sun and Yang Li. 2021. Stance detection with knowledge enhanced bert. In *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part II*, page 239–250, Berlin, Heidelberg. Springer-Verlag.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289.

Penghui Wei, Junjie Lin, and Wenji Mao. 2018. Multi-target stance detection via a dynamic memory-augmented network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1229–1232.

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*, pages 1173–1176.

Qingyang Wu, Lei Li, and Zhou Yu. 2021. Textgail: Generative adversarial imitation learning for text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14067–14075.

Chang Xu, Cecile 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)*, pages 778–783.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

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*, pages 3188–3197.