Adversarial Network with External Knowledge for Zero-Shot Stance Detection

Chunling Wang[1], Yijia Zhang[1*], Xingyu Yu[1], Guantong Liu[2], Fei Chen[1], Hongfei Lin[3]
[1]College of Information Science and Technology, Dalian Maritime University / Dalian, China
[2]College of Artificial Intelligence, Dalian Maritime University / Dalian, China
[3]College of Computer Science and Technology, Dalian University of Technology / Dalian, China
{wangchunling, zhangyijia, 1120211509, lgt, chenf}@dlmu.edu.cn
hflin@dlut.edu.cn

Abstract

Zero-shot stance detection intends to detect previously unseen targets' stances in the testing phase. However, achieving this goal can be difficult, as it requires minimizing the domain transfer between different targets, and improving the model's inference and generalization abilities. To address this challenge, we propose an adversarial network with external knowledge (ANEK) model. Specifically, we adopt adversarial learning based on pre-trained models to learn transferable knowledge from the source targets, thereby enabling the model to generalize well to unseen targets. Additionally, we incorporate sentiment information and common sense knowledge into the contextual representation to further enhance the model's understanding. Experimental results on several datasets reveal that our method achieves excellent performance, demonstrating its validity and feasibility.

Keywords: Zero-shot stance detection , Adversarial learning , External knowledge , Contrastive learning

1 Introduction

Stance detection (Küçük and Can, 2020; Mohammad et al., 2016; Augenstein et al., 2016) is a significant task in NLP, focusing on identifying the stance (e.g., against, favor, or neutral) conveyed in the text towards a given target. It can be efficiently applied to social opinion analysis (Lai et al., 2020), rumor detection (Kumar and Carley, 2019), and other research fields by mining text opinions.

Traditional intra-target stance detection (Mohammad et al., 2016) has limited applications since it requires training and testing under the same target and depends heavily on labeled data to achieve excellent performance. With the frequent and vast updates of topics on social platforms, manually labeling new targets becomes expensive and time-consuming, making it impractical to create a labeled dataset with all potential targets (Wang et al., 2020). Therefore, the study of zero-shot stance detection (Allaway and Mckeown, 2020) for unseen targets is essential and promising.

To tackle the zero-shot stance detection task, existing works generally incorporate external knowledge (Liu et al., 2021) as support for inference or introduce attention mechanisms (Allaway and Mckeown, 2020) to capture the relationships between targets, which do not explicitly model of the transferable knowledge between source and destination targets. Some methods solely focus on employing adversarial training (Allaway et al., 2021; Xie et al., 2022) to learn a target-invariant representation of the text content, disregarding the possibility that the model may encounter challenges in correctly predicting sentences that contain implicit viewpoints or require more profound understanding.

For example 1 in Table 1, the document does not explicitly mention the target "Donald Trump." If the model is unaware that Donald Trump is affiliated with the Republican Party, it is easy to misclassify the stance as neutral. Therefore, by incorporating common sense knowledge into adversarial networks and supplementing the target-related concept representations in the knowledge base, we can help the model more efficiently understand the text content, thus improving its generalization. In addition, we find a

©*Corresponding Author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

Text	Target	Gold Label
I do not understand why the **Republicans** don't dismiss him.	Donald Trump	Against
@HillaryClinton **bad** wife, **bad** role model for women, **bad** lawyer, **bad** First Lady, **bad** Senator, **horrible** Secretary of State.	Hillary Clinton	Against

Table 1. Examples of zero-shot stance detection.

certain correlation between sentiment information and stance detection (Li and Caragea, 2019). For example 2 in Table 1, when a document contains some negative words, it generally implies an Against stance. Stance detection will perform better if some sentiment knowledge can be acquired concurrently.

Motivated, on the one hand, based on the knowledge transfer ability of pre-trained models, we jointly embed the text and target into BERT and sentiment-aware BERT (noted as SentiBERT), and employ a cross-attention module to integrate the sentiment information extracted by SentiBERT with the contextual representations, resulting in semantic feature representations of the text. Meanwhile, we impose supervised contrastive learning (Liang et al., 2022) to make the model learn to distinguish stance category features in the potential distribution space. We separate the target-specific and target-invariant representations using a feature separator, then feed the target-invariant representation into the target discriminator for adversarial training, which enables the model to learn robust and transferable representations that can generalize well across different targets. On the other hand, we extract document-specific subgraphs from ConceptNet, and obtain concept representations of the common sense graph by using a graph autoencoder trained on the ConceptNet subgraph, which is fused into the text representation to enhance the model's performance. Our contributions are as follows:

(1) Our proposed ANEK model utilizes semantic information, sentiment information and common sense knowledge for zero-shot stance detection, especially adding sentiment information to assist stance detection and implicit background knowledge to enhance the model's comprehension.

(2) We employ adversarial training to learn target-invariant information to transfer knowledge effectively. Stance contrastive learning is used to enhance the inference of the model.

(3) We experimentally demonstrate that ANEK obtains competitive results on three datasets, and the extension to target stance detection is also effective.

2 Related Work

2.1 Stance Detection

Stance detection is the study of determining a text's viewpoint on a prescriptive target. (Küçük and Can, 2020). Previous studies have primarily focused on scenarios where the training and testing sets share the same target, known as intra-target stance detection (Augenstein et al., 2016; Mohammad et al., 2016). However, when new topics emerge, there is insufficient labeled data. Some studies explore cross-target stance detection (Liang et al., 2021; Wei and Mao, 2019; Xu et al., 2018), which trains a model on one target and tests it on another related target. Xu et al. (2018) presented a self-attentive model to extract shared features between targets. Wei et al. (2019) further exploited the hidden topics between targets as transferred knowledge. In contrast, zero-shot stance detection does not rely on any assumption of target correlation and is a more general study that can handle irregular target emergence.

Allaway et al. (2020) developed a dataset containing multiple targets and presented a topic-grouping attention model to capture implicit relationships between them. Liu et al. (2021) utilized the structural and semantic information of the common sense knowledge graph to enhance the model's inference. Allaway et al. (2021) regarded each target as a domain and modeled the task as a domain adaptation problem, which successfully learnd the target-invariant representation. Liang et al. (2022) designed an agent task that distinguished stance expression categories and implemented hierarchical contrastive learning. These works are considered incomplete as they overlook the impact of external knowledge containing sentiment information on the model. Whereas, we not only learn transferable target-invariant knowledge, but also take into account the introduction of multiple knowledge to enhance semantic infor-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

mation, further improving the model's predictive ability. To the best of our knowledge, we are the first to systematically introduce external knowledge into adversarial networks and achieve good results.

2.2 Adversarial Domain Adaptation

Domain adaptation mainly aims to minimize domain differences, ensure available knowledge transfer, and increase the model's generalization ability. Adversarial loss methods, inspired by the generative adversarial network (GAN) (Goodfellow et al., 2014), have been commonly applied to domain adaptation. Ganin et al. (2016) proposed a domain adversarial neural network (DANN), which utilized a gradient reversal layer to obfuscate the domain discriminator and enable the feature extractor to capture domain-invariant knowledge. Tzeng et al. (2017) presented an adversarial discriminative domain adaptation (ADDA) model, which involved a discriminative method, GAN loss, and unshared weights to decrease the domain disparity. Therefore domain adaptation is an effective solution for the zero-shot stance detection task.

2.3 External Knowledge

Neural networks enhanced with external knowledge have been used for various NLP tasks, like dialogue generation, sentiment classification, and stance detection. Ghosal et al. (2020) employed a domain adversary framework to handle cross-domain sentiment analysis and further improved the performance by injecting common sense knowledge using ConceptNet. Zhu et al. (2022) incorporated target background knowledge from Wikipedia into the stance detection model. In addition, sentiment information is useful external knowledge for stance detection tasks. Li et al. (2019) designed a sentiment classification task as an auxiliary task and built sentiment and stance vocabularies to guide attention mechanisms. Hardalov et al. (2022) adopted a pre-trained sentiment model to generate sentiment annotations for text, which improved cross-lingual stance detection performance. Based on the above work, we simultaneously consider introducing common sense and sentiment knowledge to aid stance detection.

3 Method

The structure of our ANEK model is displayed in Figure 1, which mainly contains two parts. (1) Knowledge graph training: we train a graph autoencoder using ConceptNet relation subgraphs. (2) Stance detection: we obtain context and sentiment information with pre-trained models, use contrastive learning to improve representation quality, separate features and perform adversarial learning, and finally incorporate the extracted common sense knowledge graph features to implement stance detection.

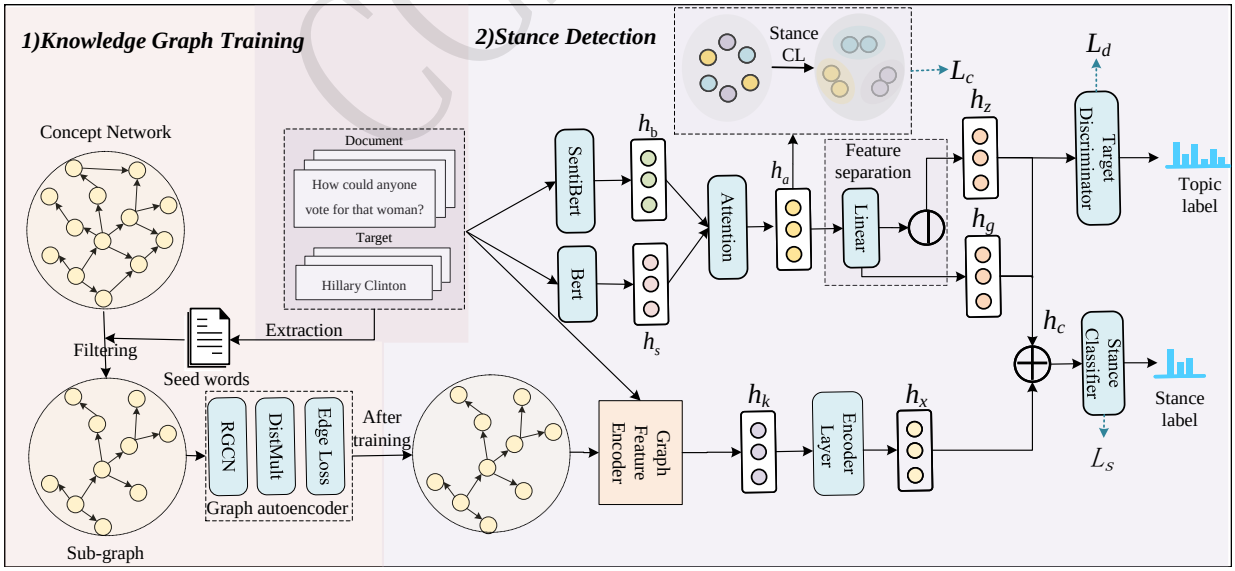

Figure 1. Overview of the ANEK model

3.1 Task Description

Suppose we are given an annotated dataset $D_s = \{x_s^i, t_s^i, y_s^i\}_{i=1}^{N_s}$ from source targets and an unlabeled dataset $D_d = \{x_d^i, t_d^i\}_{i=1}^{N_d}$ from a destination target (unknown target), where x is a document, t and y are its corresponding target and stance label, respectively, and N is the number of examples. The purpose of zero-shot stance detection is to train the model using labeled data from multiple source targets to predict the stance labels of the unknown target examples.

3.2 Knowledge Graph Training

3.2.1 Common Sense Subgraph Generation

ConceptNet is a common sense knowledge base denoted as a directed graph $G = (V, E, R)$, where concepts $v_p \in V$, edges $(v_p, r, v_q) \in E$, and $r \in R$ is the relation type of the edge between v_p and v_q. Given that ConceptNet contains tens of millions of triplet relations like (cake, IsA, dessert), we use it to construct our knowledge subgraph. To be specific, we extract unique nouns, adverbs, and adjectives from the datasets of all targets as seed words. We then extract all triples that are one edge distance away from these seed concepts to obtain a subgraph $G^{'} = \left(V^{'}, E^{'}, R^{'}\right)$.

3.2.2 Graph Autoencoder Pre-training

To integrate common sense knowledge into our model, we obtain the concept representations in the subgraph $G^{'}$ by training a graph autoencoder composed of a RGCN encoder and a DistMult decoder (Schlichtkrull et al., 2018). We feed the incomplete set of edges $\hat{E^{'}}$ from $E^{'}$ into the autoencoder. We then assigns scores to the potential edges (v_p, r, v_q) to ascertain the possibility of these edges being in $E^{'}$.

Encoder Module. To obtain enriched feature representations of the target-related concepts, we utilize two stacked RGCN encoders to compose our encoder module. RGCN can create a rich stance aggregated representation for each concept by combining related concepts in the process of neighborhood-based convolutional feature transformation. Specifically, we randomly initialize the hidden vector g_p of concept v_p and then transform it into the stance aggregated hidden vector h_p by a two-step graph convolution.

$$f(x_p, l) = \sigma(\sum_{r \in R} \sum_{q \in N_p^r} \frac{1}{a_{p,r}} W_r^{(l)} x_q + W_0^{(l)} x_p) \tag{1}$$

$$h_p = h_p^{(2)} = f(h_p^{(1)}, 2); h_p^{(1)} = f(g_p, 1) \tag{2}$$

where f denotes the encoder function with vector x_p and layer l as inputs, σ is the activation function, N_p^r indicates the neighbouring concepts of concept v_p with relation r, $a_{p,r}$ is a normalization constant, $W_r^{(l)}$, $W_0^{(l)}$ are trainable parameters.

Decoder Module. To reconstruct the edges of the graph to recover the triples' missing information, we utilize the DistMult factorization as a scoring function to calculate the score of a given triple (v_p, r, v_q).

$$s(v_p, r, v_q) = \sigma(h_p^T, R_r, h_q) \tag{3}$$

where σ is the logistic function, h_p^T is the transpose vector of concept v_p encoded by RGCN.

Training. We use negative sampling to train our graph autoencoder model (Ghosal et al., 2020). Specifically, for the triples in $\hat{E^{'}}$ (i.e., positive samples), we generate the same amount of negative examples by destroying the concepts or relation of links at random, resulting in the complete sample set Z. Our training goal is to perform binary classification between positive/negative triples with optimization using a cross-entropy loss function.

$$L_{G'} = -\frac{1}{2|\hat{E^{'}}|} \sum_{(v_p, r, v_q, y) \in Z} (y \log s(v_p, r, v_q) + (1 - y) \log(1 - s(v_p, r, v_q))) \tag{4}$$

where y is an indication that is set to 0 for negative triples and 1 for positive triples.

3.3 Stance Detection Training

3.3.1 Commonsense Feature Encoding

After training the graph autoencoder, we utilize it to generate common sense graph features for a specific target t and document x. Specifically, we extract all seed words in the document and denote them as the set K. Then the subgraph $G^{'}_{K}$ is extracted from $G^{'}$, where triples consist of concepts in K or around radius 1 of any concept in K. Next, we feed $G^{'}_{K}$ to the pre-trained RGCN encoder module and make a forward pass to get the feature representations. We calculate the average of the representations h_p for all concepts p of document x as its common sense graph features h_k. Finally, we input h_k to an encoder layer to obtain its hidden representation h_x.

$$h_x = W_x h_k + b_x \tag{5}$$

where W_x and b_x are trainable parameters.

3.3.2 Encoding with Sentiment Information

Considering that the stance of a text is influenced by sentiment information, we learn the sentiment knowledge of the text to increase prediction accuracy. Following Zhou et al. (2020), we exploit a perceptual sentiment language model (SentiBERT) to extract sentiment knowledge. We input the given document x and target t into the pretrained SentiBERT model in the form of "$[CLS]x[SEP]t[SEP]$" to obtain a hidden vector h_s with sentiment information.

$$h_s = SentiBERT([CLS]x[SEP]t[SEP]) \tag{6}$$

Moreover, to take advantage of the contextual information, we also adopt a pretrained BERT [11] model to jointly embed document x and target t to obtain a hidden vector h_b of each example.

$$h_b = BERT([CLS]x[SEP]t[SEP]) \tag{7}$$

Then h_b and h_s are concatenated, and the information of both is fused by the cross-attention module. Cross-attention can effectively capture the interdependencies between text and sentiment, facilitating the integration of knowledge and resulting in the generation of more accurate and meaningful features. The final output h_a is the hidden state of the $[CLS]$ token.

$$h_a = CrossAttention([h_b, h_s])[CLS] \tag{8}$$

3.3.3 Stance Contrastive Learning

Supervised contrastive learning can bring examples of identical categories closer together and push examples of distinct categories apart, thus learning a superior semantic representation space. To improve the generalization of the stance representation, based on the stance label information of the examples, we perform contrastive learning on their hidden vectors h_a (Liang et al., 2022). Specifically, given the hidden vectors $H = \{h_m\}_{m=1}^{N_b}$ of a batch of examples, for a specific anchor $h_m \in H$, if $h_n \in H$ and h_m have the same stance label, i.e., $y_n = y_m$, then h_n is considered to be a positive example of h_m, while other examples $h_o \in H$ are considered to be negative examples. The final contrastive loss is calculated over all positive pairs, including (h_m, h_n) and (h_n, h_m) in a batch:

$$L_c = \frac{1}{N_B} \sum_{h_m \in H} l(h_m) \tag{9}$$

$$l(h_m) = -\log \frac{\sum_{n=1}^{N_b} \mathbf{1}_{[n \neq m]} \mathbf{1}_{[y_m = y_n]} \exp(sim(\mathbf{h}_m, \mathbf{h}_n)/\tau)}{\sum_{o=1}^{N_b} \mathbf{1}_{o \neq m} \exp(sim(\mathbf{h}_m, \mathbf{h}_o)/\tau)} \tag{10}$$

$$sim(\mathbf{s}, \mathbf{t}) = \frac{\mathbf{s}^T \mathbf{t}}{||\mathbf{s}||||\mathbf{t}||} \tag{11}$$

where $1_{[m=n]} \in (0, 1)$ is an indicator function that evaluates to 1 iff $m = n$. $sim\,(\mathbf{s}, \mathbf{t})$ represents the cosine similarity of vectors $\mathbf{s}$ and $\mathbf{t}$. τ denotes a temperature parameter.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

3.3.4 Target Discriminator

The contextual representations generated by Bert and the fused sentiment information contain both target-specific and target-invariant information. Learning and exploiting transferable target knowledge is effective in enhancing the model's generalization to new targets. We separate and differentiate target-specific and target-invariant features by a simple linear transformation, which can decrease the transfer challenge with no removal of stance cues. We first extract target-specific features using a linear transformation layer (Xie et al., 2022):

$$h_g = W_g h_a + b_g \tag{12}$$

where W_g and b_g are trainable parameters. By subtracting target-specific features from h_a, the target-invariant features h_z can be obtained:

$$h_z = h_a - h_g \tag{13}$$

To further make the feature representation h_z target invariant and facilitate automatic adaptation of the model among different targets, we utilize a target discriminator to identify the target that the h_z comes from. If the discriminator cannot accurately predict the target label of h_z, we consider h_z has target-invariance. Our target discriminator is a linear network with softmax, which is trained with a cross-entropy loss function.

$$\hat{y_d} = Softmax(W_d h_z + b_d) \tag{14}$$

$$L_d = \sum_{x \in D_s} CrossEntropy(y_d, \hat{y_d}) \tag{15}$$

where W_d and b_d are the trainable parameters of the target discriminator, $\hat{y_d}$ and y_d are the predicted and true target labels. Specifically, h_z attempts to confound the target discriminator and increase the target classification loss L_d in order to learn the target-invariant features. Meanwhile, the discriminator itself struggles to decrease L_d. So we adopt the gradient reversal layer (GRL) technique, inspired by (Ganin et al., 2016), to achieve this adversarial effect by placing the GRL before the target discriminator. The essence of adversarial training is the minimum-maximum game:

$$\min_{\theta_Z} \max_{\theta_D} - \lambda \log f_D(h_z) \tag{16}$$

where θ_Z are the parameters of all network layers that generate h_z, including fine-tuned Bert, graph encoder, W_g and b_g, etc., θ_D is the discriminator parameters, and f_D is the discriminator function.

3.3.5 Stance Classifier

Since stances are essentially dependent on targets, target-specific information for each target is also indispensable. We concatenate the common sense knowledge graph features h_x, the target-invariant features h_z and the target-specific features h_g to obtain h_c, as the input for the stance classifier with softmax normalization. We minimize the stance classification loss using cross-entropy loss.

$$h_c = h_x \oplus h_z \oplus h_g \tag{17}$$

$$\hat{y} = Softmax(W_c h_c + b_c) \tag{18}$$

$$L_s = \sum_{x \in D_s} CrossEntropy(y, \hat{y}) \tag{19}$$

where W_c and b_c are the trainable parameters of the stance classifier, $\hat{y}$ and y are the predicted stance probability and ground-truth distribution.

The training goal of our proposed model is to minimize the overall loss, defined as follows:

$$L = L_s + \alpha L_c + \beta L_d \tag{20}$$

where α and β are hyperparameters.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

4 Experiments

4.1 Datasets

We conduct experiments on three publicly available datasets. 1)**SEM16** (Mohammad et al., 2016) is a Twitter dataset that contains six targets for stance detection, including the Legalization of Abortion (LA), Feminist Movement (FM), Hillary Clinton (HC), Donald Trump (DT), Atheism (A), and Climate Change is a Real Concern (CC). 2)**WT-WT** (Conforti et al., 2020) is a stance detection dataset in the financial domain. The dataset contains four targets, including ANTM_CI (AC), AET_HUM (AH), CVS_AET(CA), and CI_ESRX (CE). 3)**COVID-19** (Glandt et al., 2021) is a dataset related to COVID-19 health tasks, which includes four targets: Anthony S. Fauci, M.D. (AF), Wearing a Face Mask (WA), Keeping Schools Closed (SC), and Stay at Home (SH). Each text in the three datasets contains a stance (favor, against, neutral) for a specific target.

Following (Liang et al., 2022), we utilize the data from one target as the test set and the remaining targets as the training set. Moreover, we report the F1_avg (the Macro-averaged F1 of against and favor) as evaluation metrics.

Table 2 represents the statistics for the three datasets, listing all targets under each dataset and the number of samples labeled "favor, against, neutral, unlabeled" (where WT-WT and COVID-19 have no unlabeled samples) for each target.

Dataset	Target	Favor	Against	Neutral	Unlabeled
SEM16	DT	148	299	260	2,194
	HC	163	565	256	1,898
	FM	268	511	170	1,951
	LA	167	544	222	1,899
	A	124	464	145	1,900
	CC	135	26	203	1,900
WT-WT	CA	2,469	518	5,520	-
	CE	773	253	947	-
	AC	970	1,969	3,098	-
	AH	1,038	1,106	2,804	-
COVID-19	WA	515	220	172	-
	SC	430	102	85	-
	AF	384	266	307	-
	SH	151	201	396	-

Table 2. Statistics of the SEM16, WT-WT and COVID-19 datasets.

4.2 Experimental Implementation

We employ the pretrained SentiBERT and BERT models as the encoder, whose maximum sequence length is 85. Adam (Kingma and Ba, 2014) is used to optimize the model. In the graph autoencoder training stage, the graph batch size is 10000, the learning rate is 0.01, the dropout rate is 0.25, and we apply gradient clipping to 1.0. In the stance detection training stage, the batch size is 8, the learning rate is 1.5e-5, the dropout rate is 0.1, we train up to 50 epochs, the patience is 5, the temperature parameter for contrastive loss is 0.07. We use different seeds to train our model and record the best results.

4.3 Baselines

We compare the ANEK with several strong baselines, including **BiCond** (Augenstein et al., 2016) bidirectional conditional encoding model, **CrossNet** (Xu et al., 2018): BiCond with topic-specific attention, **TOAD** (Allaway et al., 2021): BiCond with adversarial learning, **BERT** (Kenton and Toutanova, 2019): pretrained language model, **BERT-GCN** (Liu et al., 2021): BERT with GCN for node information aggregation, **TGA Net** (Allaway and Mckeown, 2020): Bert with topic-group attention, **TPDG** (Liang

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

Model	SEM16(%)						WT-WT(%)				COVID-19(%)			
	DT	HC	FM	LA	A	CC	CA	CE	AC	AH	WA	SC	AF	SH
BiCond	30.5	32.7	40.6	34.4	31.0	15.0	56.5	52.5	64.9	63.0	30.1	33.9	26.7	19.3
CrossNet	35.6	38.3	41.7	38.5	39.7	22.8	59.1	54.5	65.1	62.3	38.2	40.0	41.3	40.4
TOAD	49.5	51.2	54.1	46.2	46.1	30.9	55.3	57.7	58.6	61.7	37.9	47.3	40.1	42.0
BERT	40.1	49.6	41.9	44.8	55.2	37.3	56.0	60.5	67.1	67.3	44.3	45.1	47.5	39.7
BERT-GCN	42.3	50.0	44.3	44.2	53.6	35.5	67.8	64.1	70.7	69.2	-	-	-	-
TPDG	47.3	50.9	53.6	46.5	48.7	32.3	66.8	65.6	74.2	73.1	48.4	**51.6**	46.0	37.3
TGA Net	40.7	49.3	46.6	45.2	52.7	36.6	65.7	63.5	69.9	68.7	-	-	-	-
PT-HCL	50.1	54.5	54.6	**50.9**	**56.5**	38.9	**73.1**	69.2	**76.7**	76.3	**58.8**	44.7	41.7	**53.3**
ANEK	**50.3**	**54.7**	**55.0**	49.0	54.1	**39.2**	71.4	**69.8**	74.8	**76.3**	52.9	49.8	**48.6**	50.3

Table 3. Experimental results on three datasets. Bold indicates the best score for each test target.

et al., 2021): GCN-based model for designing target-adaptive pragmatic dependency graphs, **PT-HCL** (Liang et al., 2022): hierarchical contrastive learning model.

4.4 Main Results

We implemented comparison experiments on three datasets and show the F1_avg results (Percentage System) in Table 3. Our proposed ANEK model presents superior performance compared to the baseline models on most target datasets. Specifically, BiCond and CrossNet perform the worst overall, as they do not consider the target invisibility to learn transferable information. Although TOAD also adopts an adversarial strategy to learn target-invariant information, its use of BiLSTM encoding is prone to poor performance in case of an unbalanced target distribution. It can be observed that it performs even less efficiently than Bert on multiple targets. As a strong baseline in NLP, BERT has good generalization because it learns rich semantic information in a large corpus, despite ignoring transferable information between targets. However, when it is applied to target transfer, it causes performance degradation due to its tendency to fit the source data. Our model explores adversarial learning based on pre-trained models, which can learn enhanced target-invariant features and improve the model's transferability.

Table 3 shows that relying solely on the introduction of common sense knowledge to help the model understand is not enough for Bert-GCN, and our model also accounts for learning sentiment information to enhance the discriminative capability of the model. We can find that ANEK slightly outperforms the PT-HCL method with hierarchical contrastive learning. Although PT-HCL obtains excellent generalization by identifying the invariant stance expressions from specific syntactic levels, it requires preprocessing the data to generate pseudo-labels, which increases the complexity of the model. Moreover, the noise brought by pseudo-labels may affect the prediction results. In contrast, our model has stronger generality and interpretability.

4.5 Ablation Study

We further designed several variants of ANEK for ablation experiments to analyze the effects of different components on the model, where "w/o CL", "w/o SK", "w/o CK", "w/o TD" denote the removal of contrastive learning, sentiment information, common sense knowledge and adversarial learning, respectively.

We report the F1_avg scores (Percentage System) of the ablation study in Table 4. The experimental results indicate that removing stance contrastive learning ("w/o CL") significantly decreases the model's performance, which suggests that we perform stance contrastive learning on the text representation assists the encoder in learning better category representations from samples, leading to better generalization. The removal of sentiment information ("w/o SK") reduces model performance, implying that the model may learn the potential relationship between stance and sentiment and make judgments with the help of sentiment knowledge. Removing common sense knowledge ("w/o CK") leads to poor performance in stance detection, indicating that introducing common sense knowledge can indeed help the model

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

understand text information and improve its reasoning ability. "w/o TD" indicates that the removal of the target discriminator becomes less effective on multiple targets, demonstrating the success of adversarial learning applied to zero-shot scenarios, generalizing to unseen targets by encouraging the encoder to generate target-invariant representations.

Model	SEM16(%)						WT-WT(%)				COVID-19(%)			
	DT	HC	FM	LA	A	CC	CA	CE	AC	AH	WA	SC	AF	SH
ANEK	**50.3**	**54.7**	**55.0**	**49.0**	**54.1**	**39.2**	**71.4**	**69.8**	**74.8**	**76.3**	**52.9**	**49.8**	**48.6**	**50.3**
w/o LC	49.2	52.8	52.9	47.8	53.2	38.0	69.2	66.5	73.2	75.2	51.3	48.2	48.1	49.2
w/o SK	48.7	51.8	53.4	47.2	52.0	37.8	68.1	67.5	71.3	74.0	51.0	49.3	47.2	48.0
w/o CK	48.0	52.4	53.0	46.8	51.1	36.5	67.6	66.8	72.0	73.8	49.7	48.7	46.5	47.9
w/o TD	47.8	51.2	52.3	46.5	52.9	37.8	69.0	68.8	72.6	73.3	50.4	47.9	47.8	47.2

Table 4. Experimental results of the ablation study.

4.6 Generalizability Analysis

We further performed experiments on the SEM16 dataset for cross-target stance detection and report the F1_avg results (Percentage System) in Table 5. The cross-target stance detection task is treated as a particular zero-shot setting, as we need to train using data from a source target related to the test target. Table 5 illustrates that our ANEK model achieves better performance. We can also find that the cross-target setting outperforms the zero-shot setting, which indicates that knowing the relationship between targets in advance can learn more reliable target-invariant representations to generalize to unseen targets, illustrating the challenges of zero-shot stance detection. Additionally, enhancing the understanding and generalization of the model by introducing external knowledge is also effective.

Model	SEM16(%)			
	FM→LA	LA→FM	HC→DT	DT→HC
BiCond	45.0	41.6	29.7	35.8
CrossNet	45.4	43.3	43.1	36.2
BERT	47.9	33.9	43.6	36.5
TPDG	58.3	54.1	50.4	52.9
PT-HCL	**59.3**	54.6	53.7	55.3
ANEK	58.5	**54.8**	**54.3**	**56.4**

Table 5. Experimental results of cross-target stance detection. "FM→LA" indicates training on FM, testing on LA, etc.

Text	Target	Gold Label	BERT	TOAD	ANEK
Your have to wonder if Hillary will attempt to replace #ObamaCare with #HillaryCare.	Donald Trump	Against	Neutral	Against	Against
Donald trump is way better than ANY candidate out there. Because he's real, not a lobbyist backed puppet.	Donald Trump	Against	Favor	Favor	Against
I do not understand why the Republicans don't dismiss him.	Donald Trump	Against	Neutral	Neutral	Against
.....and some, I assume, are good people.	Donald Trump	Against	Favor	Favor	Favor

Table 6. Four cases of the predictions by BERT, TOAD and ANEK.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

4.7 Case Study

To qualitatively analyze our model, we conduct a case study and error analysis. We select four cases from the test data of SEM16 and compare our results to the predictions of BERT and TOAD. Table 6 reports these results. In the first case, our model and TOAD with adversarial learning output the correct labels, while the output of BERT is wrong. We believe that because the training data contains the target "Hillary Clinton," the model learns the election relationship between the two targets and transfers the knowledge, and semantically focuses more on the stance-related words rather than the target words, with a robust target generalization. In the second case, only our method makes the correct prediction, demonstrating that depending only on contextual information is insufficient. Adding sentiment information strengthens the model's comprehension of texts with a sarcastic sentiment. In the third case, our method still correctly predicts the outcome. Although no words about Trump appear in the text, we speculate that the model learns the hidden connection between "Republican" and "Donald Trump" and understands the implied meaning of the text, further confirming the validity of common sense knowledge.

In the fourth case, all models output incorrect results. We suspect that this is because the text is too brief, resulting in less valid information being learned, and the background knowledge is too complex, which reveals that we can explore data augmentation methods in the future to improve the performance of zero-shot stance detection by expanding the data.

5 Conclusion

This paper proposes an adversarial network with external knowledge (ANEK) to handle the zero-shot stance detection task. The model applies adversarial learning based on pre-trained models to ensure knowledge transferability, and introduces common sense knowledge and sentiment information to enhance the model's deep understanding and assist stance detection. In addition, stance contrastive learning is used to improve the model's generalization. The experimental results on three benchmark datasets indicate that our method performs competitively on some unseen targets. In future work, we will design a data enhancement method to alleviate the data scarcity problem in zero-shot settings and improve performance.

Acknowledgements

This work is supported by a grant from the Social and Science Foundation of Liaoning Province (No. L20BTQ008)

References

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.

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*, pages 1715–1724.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. Kingdom: Knowledge-guided domain adaptation for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. 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)*, pages 1596–1611.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks (2014). *arXiv preprint arXiv:1406.2661*, 1406.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-shot cross-lingual stance detection with sentiment-based pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10729–10737.

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. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Sumeet Kumar and Kathleen M Carley. 2019. Tree lstms with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5047–5058.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

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)*, pages 6299–6305.

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.

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*, pages 2738–2747.

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.

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.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176.

Zhen Wang, Qiansheng Wang, Chengguo Lv, Xue Cao, and Guohong Fu. 2020. Unseen target stance detection with adversarial domain generalization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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.

Feng Xie, Zhong Zhang, Xuechen Zhao, Jiaying Zou, Bin Zhou, and Yusong Tan. 2022. Adversarial learning-based stance classifier for covid-19-related health policies. *arXiv e-prints*, pages arXiv–2209.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

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.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics*, pages 568–579.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2070–2075.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.