C-STANCE: A Large Dataset for Chinese Zero-Shot Stance Detection

Chenye Zhao[♣] **Yingjie Li**[♡] **Cornelia Caragea**[♣]
[♣] University of Illinois at Chicago
[♡] Westlake University
{czhao43,cornelia}@uic.edu
liyingjie@westlake.edu.cn

Abstract

Zero-shot stance detection (ZSSD) aims to determine whether the author of a text is in favor of, against, or neutral toward a target that is unseen during training. Despite the growing attention on ZSSD, most recent advances in this task are limited to English and do not pay much attention to other languages such as Chinese. To support ZSSD research, in this paper, we present C-STANCE that, to our knowledge, is the first Chinese dataset for zero-shot stance detection. We introduce two challenging subtasks for ZSSD: target-based ZSSD and domain-based ZSSD. Our dataset includes both noun-phrase targets and claim targets, covering a wide range of domains. We provide a detailed description and analysis of our dataset. To establish results on C-STANCE, we report performance scores using state-of-the-art deep learning models. We publicly release our dataset and code to facilitate future research.[1]

1 Introduction

Stance detection aims to automatically predict whether the author of a text is in favor of, against, or neutral toward *a specific target* (Mohammad et al., 2016b; Küçük and Can, 2020; ALDayel and Magdy, 2021), e.g., epidemic prevention, gasoline price, or equal rights. The stance can provide useful information for important events such as policy-making and presidential elections.

Early works focus on two types of stance detection tasks: in-target stance detection, where classifiers are trained and tested on data from the same set of targets (Hasan and Ng, 2014; Mohammad et al., 2016b; Graells-Garrido et al., 2020) and cross-target stance detection, where classifiers are trained on source targets that are related to destination targets (Augenstein et al., 2016; Wei and Mao, 2019), but destination targets are unseen during training. However, it is impractical to include all possible or related targets in the training set. More recently, zero-shot stance detection (ZSSD) has been identified as a promising direction (Allaway and McKeown, 2020) to evaluate classifiers on a large number of unseen (and unrelated) targets. ZSSD is more similar to the situations in practice and has received a lot of attention (Liu et al., 2021; Luo et al., 2022; Liang et al., 2022b).

Despite the growing interest in stance detection, the task has several limitations. First, most recent advances in stance detection are limited to English (Mohammad et al., 2016b; Allaway and McKeown, 2020; Conforti et al., 2020b; Li et al., 2021a; Glandt et al., 2021), and pay little attention to other languages such as Chinese (Xu et al., 2016) although large amounts of online data with expressions of stance are available in other languages. Second, the current ZSSD task (Allaway and McKeown, 2020) aims to detect the stance of unseen targets. However, these unseen targets come from the same domain of training targets with similar meanings, which makes the task less challenging. Third, current stance detection datasets include targets either as noun phrases (Mohammad et al., 2016b; Glandt et al., 2021) or as claims (Ferreira and Vlachos, 2016; Derczynski et al., 2017). However, in practice, stance is usually expressed toward both noun phrases and claims. Models trained only on noun-phrase targets do not necessarily work well for claim targets and vice versa. Little attention is paid toward incorporating targets of both types.

In an effort to minimize these drawbacks, we present C-STANCE, the first Chinese zero-shot stance detection dataset. Our dataset is collected from Sina Weibo, one of the most popular Chinese social media sites (akin to Twitter). We consider two practical scenarios for zero-shot stance detection, i.e., target-based and domain-based ZSSD. **Subtask A: target-based zero-shot stance detection**. Subtask A is similar to the previous ZSSD task, where stance detection classifiers are evalu-

[1] https://github.com/chenyez/C-STANCE

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13369–13385

Microblog	请赶紧打疫苗，勤洗手把口罩戴起来。随着新冠病例的剧增，医疗资源不足必然引起医患矛盾的爆发，也请大家能互相体谅。 Please get vaccinated quickly, wash hands frequently and put on your mask. With the sharp increase in Covid-19 cases, the shortage of medical resources will inevitably lead to the outbreak of conflicts between doctors and patients. Please understand each other.
Noun-phrase target/Stance	1. 新冠疫苗 Covid-19 vaccine / Favor 2. 医患矛盾 Conflict between doctors and patients / Against
Claim target/ Stance	1. 还是要去打疫苗，做好自我防护，尽量别阳，不去挤占医疗资源。 We should still get vaccinated and do self-protection, try not get covid and not to take medical resources. / Favor 2. 那些戴好口罩，勤消毒的人都阳了，所以说防护没啥用的。 Those who wear masks and disinfect frequently still get covid, so it is useless to defend. / Against 3. 免疫力真的很重要，平时就要加强自身锻炼，增强免疫力。 Immunity is really important, we need to strengthen our own exercise at ordinary times to enhance immunity / Neutral

Table 1: Examples of noun-phrase targets and claim targets for a microblog in the "**Covid Epidemic**" domain of our C-STANCE dataset.

ated using a large number of completely unseen targets. **Subtask B: domain-based zero-shot stance detection**. Subtask B is our newly proposed ZSSD task where stance detection classifiers are evaluated using a large number of unseen targets from completely new domains. Additionally, C-STANCE captures a more diverse set of targets including both noun-phrase targets and claim targets compared with existing datasets. An example from our dataset is shown in Table 1. As we can see from the table, the author of the microblog is in favor of the noun-phrase target "Covid-19 vaccine" and against "the conflict between doctors and patients". The author also opposes claim target 2, whose main idea is to deny the necessity of self-protection.

Our contributions can be summarized as follows: 1) We present C-STANCE, the first large Chinese zero-shot stance detection dataset. Our dataset is composed of 48,126 annotated text-target pairs. C-STANCE is more than 2.5 times larger than the English ZSSD VAST dataset (Allaway and McKeown, 2020) and more than 16 times larger than the existing Chinese stance detection dataset by Xu et al. (2016). We provide detailed description and analysis of our dataset; 2) We include two challenging ZSSD subtasks: target-based zero-shot stance detection and domain-based zero-shot stance detection for C-STANCE; 3) We consider a more diverse set of targets including both noun phrases and claims in C-STANCE as well as multiple targets per input text (see Table 1); 4) We establish baseline results using both traditional models and pre-trained language models and show that C-STANCE is a challenging new benchmark. For example, our best-performing model based on RoBERTa achieves only 78.5% $F1_{macro}$ for subtask A.

2 Related Work

Most previous stance detection datasets are constructed for the English language (Mohammad et al., 2016b; Conforti et al., 2020b; Glandt et al., 2021). Particularly, VAST is the only dataset for zero-shot stance detection. Even though recent years have witnessed an emerging trend of constructing stance detection datasets of other languages (Xu et al., 2016; Taulé et al., 2017; Swami et al., 2018; Lai et al., 2020; Vamvas and Sennrich, 2020), Chinese stance detection datasets are still very scarce. Xu et al. (2016) developed the first Chinese stance dataset. The dataset focuses on in-target stance detection and only includes 3,000 examples from 6 targets. In contrast, we propose the first large-scale Chinese dataset for zero-shot stance detection. Our C-STANCE which includes 48,126 samples with 11,623 noun-phrase targets and 28,581 claim targets enables multiple stance detection tasks and covers a wide range of domains.

Besides classifying stance detection by target type (noun phrases or claims), we can also categorize the task as in-target, cross-target, and zero-shot stance detection. Most previous works focused on in-target stance detection where a classifier is trained and evaluated on the same target (Zarrella and Marsh, 2016; Wei et al., 2016; Vijayaraghavan et al., 2016; Mohammad et al., 2016b; Du et al., 2017; Sun et al., 2018; Wei et al., 2018; Li and Caragea, 2019, 2021b). However, it is usually hard to obtain sufficient annotated data for each target and conventional models perform poorly when generalized to data of unseen targets. This motivated the research on cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Wei and Mao, 2019; Zhang et al., 2020; Li et al., 2021b), where a classifier is adapted from different but related targets. However, cross-target stance detection still requires prior human knowledge of the destination target and how it is related to the training targets. Thus models developed for cross-target stance detection are still limited in their capability to generalize to a wide range of unseen targets. Zero-shot

Authors	Source	# Target(s)	Target Type	Language	Size
Ferreira and Vlachos (2016)	News articles	300	Claim	English	2,595
Derczynski et al. (2017)	Twitter	305	Claim	English	5,568
Gorrell et al. (2019)	Twitter, Reddit	8,574	Claim	English	8,574
Vamvas and Sennrich (2020)	Political Comments	194	Claim	English, French, Germany, Italian	67,000
Xu et al. (2016)	Weibo	7	Noun-phrase	Chinese	5,000
Mohammad et al. (2016b)	Twitter	6	Noun-phrase	English	4,870
Swami et al. (2018)	Twitter	1	Noun-phrase	English, Hindi	3,545
Conforti et al. (2020b)	Twitter	5	Noun-phrase	English	51,284
Allaway and McKeown (2020)	News Comments	5,634	Noun-phrase	Egnlish	18,545
Glandt et al. (2021)	Twitter	4	Noun-phrase	English	6,133
Li et al. (2021a)	Twitter	3	Noun-phrase	English	21,574
Lai et al. (2020)	Twitter	6	Noun-phrase	English, Spanish, Catalonia, French, Italian	14,440
C-STANCE (ours)	Weibo	40,204	Noun-phrase, Claim	Chinese	48,126

Table 2: Comparison of stance detection datasets.

stance detection (ZSSD) which aims to identify the stance toward a large number of unseen targets has attracted considerable attention in recent years. Allaway and McKeown (2020) developed a dataset for ZSSD which is called VAried Stance Topics (VAST) that includes thousands of targets. Based on VAST, many ZSSD models have been developed (Liu et al., 2021; Liang et al., 2022a,b; Luo et al., 2022; Li et al., 2023). In contrast to VAST, we include two types of ZSSD subtasks in C-STANCE. The first subtask is the target-based ZSSD which is similar to the VAST setting. The second subtask is the domain-based ZSSD where classifiers are evaluated on unseen targets from completely new domains, which is a more challenging task.

Target-specific stance detection is the most common stance detection task (ALDayel and Magdy, 2021), which aims to predict the stance label toward a target, which could be a figure or controversial topic (Hasan and Ng, 2014; Mohammad et al., 2016a; Zotova et al., 2020; Conforti et al., 2020a,b). Multi-target stance detection is another type of stance detection task that aims to jointly identify the stance toward two or more targets in the same text (Sobhani et al., 2017; Darwish et al., 2017; Li and Caragea, 2021a). Unlike target-specific and multi-target stance detection where targets are usually noun phrases (phrase-based stance detection), claim-based stance detection aims to predict the stance toward a specific claim, which could be an article headline or a rumor's post (Qazvinian et al., 2011; Derczynski et al., 2015; Ferreira and Vlachos, 2016; Bar-Haim et al., 2017; Derczynski et al., 2017; Gorrell et al., 2019). However, less attention has been paid to incorporating both noun-phrase targets and claim targets into one dataset. Comparatively, our dataset supports data for both claim-based stance detection and phrase-based stance detection as well as captures multiple targets per input text (see examples from our dataset in Appendix A). We compare our C-STANCE dataset with previous stance detection datasets in Table 2.

3 Dataset Construction

In this section, we describe the creation and particularities of C-STANCE, a large comprehensive stance detection dataset composed of 48,126 annotated instances covering a wide range of domains.

3.1 Data Collection

We collect microblogs using the Weibo API from July 26th, 2022, to November 16th, 2022. Similar to prior works (Mohammad et al., 2016b; Glandt et al., 2021; Li et al., 2021a), our crawling is performed using query keywords. To cover a wide range of domains on Weibo, we start by using the domain names listed on the *Weibo hot list* page as keywords for crawling (e.g., society, education, etc.). After we get our initial set, we select the most frequent words as supplementary keywords for the next round of crawling to gradually expand our keyword set. The full list of keywords that were used is provided in Appendix B. We end up collecting 60,000 microblogs.

3.2 Keywords Selection

After data collection, we filter keywords that are most suitable for the task of stance detection. We perform the following steps for keyword filtering: 1) We manually detect and remove keywords that often contain advertising content (e.g., beauty, renting, motor show, etc.), which are not suitable for stance detection as the purpose of those microblogs

Domain		Query Keywords
新冠疫情 Covid Epidemic	CoE	防疫 epidemic prevention, 封控 sealed management,口罩 mask, 群体免疫 herd immunity, 居家办公 work-from-home, 疫苗 vaccine 新冠共存 co-existence with coronavirus, 加强针 booster
世界事件 World Events	WE	世界新闻 world news, 乌克兰 Ukraine, 俄罗斯 Russia, 移民 migrant, 人口负增长 negative population growth, 战争 war, 选举 election, 大选 general election
文化教育 Cultural and Education	CuE	素质教育 quality education, 鸡娃 force kids to compete, 文化输出 cultural output, 传统文化 traditional culture 公立教育 public education, 流行文化 pop culture
娱乐消费 Entertainment and Consumption	EC	物价 prices, 油价 gasoline price, 直播带货 livestream shopping, 短视频 short video, 保险 insurance, 消费观 consumption concept, 微商 wechat business, 苹果手机 iphone, 股市 stock market, 媒体 media
体育 Sports	S	世界杯 World Cup, NBA, 男足 men's football, 女足 women's football, 体育 sports
权利 Rights	R	性别平等 gender equality, 女权 women's rights, 性少数群体 LGBTQ, 医患 doctors and patients, 平权 equal rights
环保 Environmental Protection	EP	气候变化 climate change, 垃圾分类 garbage classification, 环保意识 environmental awareness, 新能源 new energy

Table 3: The domains used in our dataset and the selected query keywords for each domain.

	Noun-phrase targets			Claim targets		
Domain	Favor	Against	Neutral	Favor	Against	Neutral
CoE	1,247	1,444	783	1,782	1,782	1,782
WE	641	870	1,616	1,590	1,590	1,590
CuE	1,108	734	554	1,206	1,206	1,206
EC	1,480	1,355	1,175	2,051	2,051	2,051
S	766	435	885	1,059	1,059	1,059
R	940	1,020	532	1,276	1,276	1,276
EP	633	264	556	732	732	732
Overall	6,815	6,122	6,101	9,696	9,696	9,696

Table 4: Label distribution for noun-phrase targets and claim targets in each domain.

is not to discuss controversial topics but to promote the sales of particular products. 2) For stance detection, we show more interest in controversial topics and keywords where people may express different stances (favor, against, or neutral) toward targets related to these keywords. Otherwise, models would predict the stances based on keywords information instead of the contents of microblogs and the targets. Therefore, we filter out keywords that people tend to show uni-stances on, e.g., poverty, delicious food, traveling, camera, etc., and keywords where microblogs often express personal feelings (e.g., "my girlfriend", "my mood", etc). After this filtering step, we select 45 keywords that cover controversial topics. We summarize the 45 keywords into 7 domains: "Covid Epidemic" (CoE), "World Events" (WE), "Cultural and Education" (CuE), "Entertainment and Consumption" (EC), "Sports" (S), "Rights" (R), and "Environmental Protection" (EP) which can be seen in Table 3.

3.3 Preprocessing

We perform several preprocessing steps to ensure the quality of our dataset. 1) We remove microblogs with less than 50 or more than 200 words. From our observations, microblogs with less than 50 words usually are either too noisy or cannot cover enough information to express stances toward multiple targets. Microblogs with more than 200 words are usually technical articles that contain little stance-related discussion. 2) We remove duplicates and reposted microblogs. 3) We keep only microblogs in Chinese. We leave the multilingual dataset as our future work. 4) We manually identify a set of phrase lexicon for advertisements (e.g., check the link below, follow our WeChat public account, scan the QR code, click to join, etc.). We filter out all microblogs containing phrases in this lexicon. 5) We remove the emojis, URLs in microblogs as they may introduce noise to the dataset. After preprocessing, our corpus reduces to around 25,000 examples. We randomly sample around 215 microblogs for each of the 45 keywords, obtaining 9,696 microblogs for annotation.

3.4 Data Annotation

We gather annotations using Taojinniwo,[2] a Chinese crowd-sourcing company that provides annota-

[2]http://sjbz.itaojin.cn/

		# Examples		# Targets		Avg. Length			
		N	C	N	C	N	C	MB	# Unique MB
Subtask A	Train	13,258	20,160	6,093	19,694	3.7	25.7	101.9	6,740
	Val	2,865	4,419	2,665	4,400	4.6	26.3	104.0	1,473
	Test	2,915	4,509	2,865	4,487	4.7	26.5	105.7	1,503
Subtask B (Covid Epidemic)	Train	12,379	18,984	7,519	18,585	4.0	26.0	102.4	6,690
	Val	2,249	3,447	2,208	3,436	4.6	26.0	104.8	1,087
	Test	3,474	5,346	1,896	5,211	3.7	25.7	103.6	1,786

Table 5: Dataset split statistics for subtask A and subtask B ("Covid Epidemic" as the zero-shot domain). N, C, MB represent noun-phrase targets, claim targets, and microblogs, respectively.

tion services for big AI companies (e.g., Baidu, JD, etc.). To ensure the annotation quality, we employ strict requirements for annotators: 1) Annotators should reside in China; 2) Annotators should have college degrees. Moreover, we randomly select 10% of each annotator's annotations for quality checks. If an annotator has an acceptance rate of less than 90%, we discard their annotations completely and re-send them for labeling using other qualified annotators. We annotated data for noun-phrase targets and for claim targets as detailed below. The label distribution for each domain is shown in Table 4.

3.4.1 Annotation for Noun-Phrase Targets

The annotation for noun-phrase targets is performed in two steps. In step 1, one annotator is asked to detect at least 2 targets from each microblog. Annotators are given the following instructions: *"You should identify 2 or more targets in the form of noun phrases. Targets should satisfy the following requirements: 1) Targets should be the main focus of the microblog instead of the trivial details; 2) Targets should be public topics that people may take stances on; 3) Avoid selecting targets to which most people may express the same stance, e.g., illegal charge."*. In step 2, we ask three annotators to assign a stance label to each microblog-target pair. The instructions are given below: *"Based on the message that you learned from the microblog, predict the stance that the author would take for the given target as "Favor", "Against", or "Neutral"*. We take the majority vote among stance annotations from the three annotators to obtain stance labels. For 9,696 microblogs, we collected 19,038 annotated instances (around 2 targets per instance). The inter-annotator agreement measured by Krippendorff's alpha (Krippendorff, 2011) is 0.60, and a percentage agreement of 75%. We see that while the task is challenging, annotators agree the majority of the time. We can observe from Table 4 that the "World Events" (WE) domain and the "Sports" (S) domain have the highest percentage in the "Neutral" class. This might be because these domains include more microblogs related to news. Moreover, people are showing a higher percentage of "Against" stances toward targets in the "Covid Epidemic" (CoE) and "Rights" (R) domains, where more contrary opinions are often expressed.

3.4.2 Annotation for Claim Targets

The goal of this annotation task is to identify three claims, to which the microblog takes favor, against, and neutral stances, respectively. Annotators are provided with the following instructions: *"After reading the microblog, write the following three claims: 1) The author is definitely in favor of the point or message of the claim (favor); 2) The author is definitely against the point or message of the claim (against); 3) Based solely on the microblog content, we cannot know whether the author supports or is against the point or message of the claim (neutral)."*. To pose challenges to the ZSSD task, we have some additional requirements: First, claims with label favor should not be a direct copy of the microblog content. Second, claims with labels against should not be the simple negation of the microblog content (e.g., adding "not" before verbs). Models may easily detect such language patterns and predict the stance without considering the content of microblog-claim pairs.

Note that our claim annotation differs from the task of rumor detection (Zubiaga et al., 2015; Derczynski et al., 2017), where claims are replies stemming from the text. Some of such claims may miss information mentioned in the text (e.g., Text: Coronavirus is made by the alien. Claim: I don't believe that.). Our task focuses on predicting the stance toward a claim that discusses the same topic and does not omit any necessary information (e.g., Claim: I believe the Coronavirus is made by some terrorists). We collect 29,088 annotated microblog-claim pairs. For quality assurance, we hide the stance label and

ask another group of annotators to annotate the stance label for a subset of microblog-claim target pairs. The two annotation groups agree on 95% of the annotation. We observe from Table 4 that each domain has a balanced label distribution. This is because we annotate one claim for each stance label from each microblog.

3.5 Dataset Split

We split the annotated data into training, validation, and test sets for the target-based ZSSD (subtask A) and the domain-based ZSSD (subtask B). For subtask A, we separate the dataset following the VAST dataset (Allaway and McKeown, 2020): the training, validation, and test sets do not share any microblogs and targets with each other. We randomly select 70% of unique annotated microblogs for the training set and split the remainder evenly for the validation and the test set. The dataset distribution is shown in Table 5. We have 2,865 unique zero-shot noun-phrase targets and 4,487 unique zero-shot claims for 1,503 unique microblogs in the test set, with the average length of 4.7, 26.5, and 105.9 for noun-phrase targets, claim targets, and microblogs, respectively. We also report the average percentage of tokens in targets that overlap with tokens in microblogs (see Appendix C).

For subtask B, we use the data from six domains (source) for training and validation, and the data from the left-out domain (zero-shot) as the test set. In the end, we have 7 dataset splits for subtask B with one dataset split for each of the 7 domains where each domain in turn is used as the test set. To ensure there are no overlapping targets between the source domains and the zero-shot domain, we remove data with overlapping targets from the source domains in each split. We then split the source domains into the training and the validation set without overlapping microblogs and targets. The statistics when using the "Covid Epidemic" as the zero-shot domain are shown in Table 5. The full statistics of subtask B are shown in Appendix D.

Because of the linguistic variations in the noun-phrase target expressions, we study the prevalence of *LexSimTopics* (Allaway and McKeown, 2020) between the training and the test set, which is defined as test targets that have more than 0.9 cosine similarities with any train targets in the word embedding space (Bojanowski et al., 2017). We observe that for subtask A, we have 11% *LexSimTopics* in the test set. Whereas for the "Covid Epidemic" domain in subtask B, we only have 7% *LexSimTopics*. This implies that subtask B is more challenging as the training and test targets are more different from each other. Comparatively, VAST dataset has 16% *LexSimTopics* in the zero-shot test set which is higher than our task.

4 Experimental Settings

In this section, we introduce the baselines in Section 4.1 and the training settings in Section 4.2.

4.1 Baseline Methods

To evaluate C-STANCE, we run experiments with the following baselines. **BiCE** (Augenstein et al., 2016) and **CrossNet** (Xu et al., 2018) predict the class label using the conditional encoding of BiLSTM models. **TGA-Net** (Allaway and McKeown, 2020) implicitly captures relationships between targets using generalized topic representations to assist stance classification. We also consider the base version of **BERT** (Devlin et al., 2019) trained using the whole word masking (wwm) on Chinese Wikipedia (Cui et al., 2020), the 12-layer **RoBERTa** (Liu et al., 2019) and **XLNet** (Yang et al., 2019) pre-trained on Chinese news, Q&A, and BaiduBaike (Cui et al., 2020).

4.2 Training Settings

We perform experiments using an NVIDIA RTX A5000 GPU. Our experiments are conducted based on PyTorch (Paszke et al., 2019). The validation set was used to determine the hyperparameters for the models. For BiCE and CrossNet, we used the AdamW (Loshchilov and Hutter, 2019) with a learning rate of 0.001. Each model was trained for 20 epochs, with a mini-batch size of 64. For TGA-Net, we followed hyperparameters suggested in the previous work (Allaway and McKeown, 2020). For BERT, RoBERTa, and XLNet, we used the AdamW with a learning rate of 5e-6. Models were fine-tuned for 5 epochs using a mini-batch size of 32. The total training time is less than 3 hours.

5 Results

In this section, we first perform experiments on subtask A and subtask B. We then conduct experiments on cross-lingual stance detection using C-STANCE and the previous English ZSSD VAST dataset. We also study the impact of incorporating both noun-phrase targets and claims targets into one dataset. Lastly, we perform the spuriosity analysis for claim

	Mixed targets				Noun-phrase targets				Claim targets			
	Con	Pro	Neu	All	Con	Pro	Neu	All	Con	Pro	Neu	All
BiCE	.490	.408	.443	.447	.560	.515	.590	.555	.335	.358	.302	.332
Cross-Net	.526	.541	.592	.553	.607	.567	.601	.592	.441	.395	.588	.475
TGA Net	.565	.599	.637	.600	.694	.674	**.670**	.679	.488	.625	.699	.604
BERT	.758	.763	.798	.773	.708	.693	.647	.683	.797	.827	.899	.841
RoBERTa	**.775**	**.769**	**.811**	**.785**	.712	**.701**	.669	.694	.797	.819	.899	.838
XLNet	.767	**.769**	.804	.780	**.721**	**.701**	.667	**.696**	**.805**	**.829**	**.900**	**.845**

Table 6: Comparison of different models on C-STANCE subtask A. The performance is reported using F1 score for the against (Con), favor (Pro), neutral (Neu), and the $F1_{macro}$ (All). Reported results are averaged over four runs.

targets. Each result is the average of 4 runs with different initializations. Similar to prior works (Allaway and McKeown, 2020; Liang et al., 2022b), we use the F1 for each class and the macro-averaged F1 of all classes as evaluation metrics.

5.1 Target-based Zero-Shot Stance Detection

Target-based zero-shot stance detection (subtask A) aims to evaluate the classifier on a large number of completely unseen targets (Allaway and McKeown, 2020). Our experiments are performed using the full dataset with mixed targets (both noun phrases and claims), the dataset with noun-phrase targets, and the dataset with claim targets, respectively.

Experimental results are shown in Table 6. First, we can observe that transformer-based models show better performance than RNN-based models, demonstrating the effectiveness of the pre-trained transformer models. Moreover, RoBERTa and XLNet outperform BERT in most metrics, suggesting the effectiveness of additional training performed by RoBERTa and XLNet to address different limitations of BERT. Second, transformer-based models perform better on claim targets than noun-phrase targets. This might be because transformer models are better at capturing contextual information and claims are usually composed of more contextual information than noun phrases. Comparatively, BiCE and CrossNet perform worse on claim targets than noun-phrase targets, showing that claim targets are more challenging for RNN-based models. We also notice that TGA-Net achieves worse performance on claim targets. This might be because the model requires clustering based on target representations, which is more difficult for the claim targets. Third, model performance for the mixed target is higher than the noun-phrase targets and lower than the claim targets. This suggests that ZSSD models that can properly utilize both types of targets are still needed, which we leave as our future work.

Model	Data	CoE	WE	CuE	EC	S	R	EP
BiCE	M	.347	.413	.376	.393	.413	.360	.400
	N	.447	.546	.479	.506	.539	.459	.493
	C	.305	.296	.289	.304	.313	.304	.286
CrossNet	M	.374	.375	.370	.392	.374	.351	.386
	N	.489	.582	.497	.523	.530	.471	.522
	C	.243	.260	.308	.260	.279	.244	.253
TGA-Net	M	.570	.581	.598	.598	.609	.608	.592
	N	.577	.667	.632	.629	.654	.619	.642
	C	.584	.585	.598	.608	.613	.603	.615
BERT	M	.753	.773	.768	.762	.775	.772	.777
	N	.594	.664	.641	.641	.671	.621	.647
	C	.828	.835	.836	.824	.841	.832	.866
RoBERTa	M	.755	**.776**	**.779**	**.774**	**.785**	**.784**	**.795**
	N	**.602**	.676	.647	**.655**	**.687**	.635	**.670**
	C	.822	.833	.834	.820	.842	**.836**	**.879**
XLNet	M	**.758**	.763	.778	.767	.777	.777	.781
	N	.594	**.680**	**.657**	.652	.674	**.640**	.654
	C	**.830**	**.839**	**.840**	**.832**	**.845**	.834	.874

Table 7: Comparison of $F1_{macro}$ of different models on C-STANCE subtask B. Models are trained and evaluated using datasets for 7 zero-shot domain settings. Results are averaged over four runs.

5.2 Domain-based Zero-Shot Stance Detection

Domain-based stance detection (subtask B) focuses on evaluating classifiers using unseen topics from completely new domains. Particularly, we select one domain as the zero-shot domain and the rest six domains as source domains. We train and validate models using data from source domains and test models using data from the zero-shot domain. We have seven zero-shot domain settings (each with a different zero-shot domain). Similar to subtask A, our experiments are performed using the full dataset with mixed targets, data with noun-phrase targets, and data with claim targets, respectively.

Results are shown in Table 7. First, we can observe that among the seven zero-shot domain settings, most models show the highest performance when predicting stances for claim targets and the mixed targets from the "Environmental Protection" domain. For example, RoBERTa achieves the highest $F1_{macro}$ of 0.879 for the claim targets, improving its performance over the rest domains by up to 5.9%. Second, stances for noun-phrase targets

Train/Val	Test	Con	Pro	Neu	All
C	V	.431	.434	.424	.430
C	V-MT	.386	.412	.372	.390
V	C	.461	.483	.142	.362
V	C-MT	.356	.436	.121	.304

Table 8: Cross-lingual ZSSD performance of mBERT using VAST and C-STANCE (denoted as V and C, respectively). "MT" represents machine translation.

from the "Sports" and the "World Events" domains are easier to predict than the other domains, where RoBERTa and XLNet achieve the highest $F1_{macro}$ of 0.687 and 0.680, respectively. This might be because sports and world events are domains with a higher percentage of microblogs discussing news, which usually captures more diverse target ranges than the other domains. Moreover, we also observe that in most cases, the "Covid Epidemic" is the most difficult domain to predict for all targets, suggesting that the "Covid Epidemic" domain shares the least domain knowledge with the other domains, making it the most difficult zero-shot domain for domain-based ZSSD.

For mixed targets experiments, we also report the results for test sets of only noun-phrase targets and only claim targets separately in Appendix E (i.e., training on mixed targets and testing on noun-phrase targets or training on mixed targets and testing on claim targets).

5.3 Cross-Lingual Zero-Shot Stance Detection

To better understand the difference between the existing English ZSSD dataset (VAST) and our Chinese C-STANCE dataset, we perform experiments on cross-lingual zero-shot stance detection between the two datasets. Particularly, we fine-tune a multilingual transformer model BERT (mBERT) (Devlin et al., 2019). The model is pre-trained on 104 languages. We train and validate mBERT using one dataset, and test the model using the other dataset. During the test stage, we experiment with both the original test set and the test set translated into the other language using Google Translate [3].

As shown in Table 8, models trained on VAST perform poorly on the neutral class for the C-STANCE, while models trained on C-STANCE show much higher performance. The results imply that the neutral class in C-STANCE is more challenging than VAST. This is because data for the neutral class in VAST is generated by randomly permuting existing targets and texts, which may generate easy-to-detect text-target pairs. Comparatively, for C-STANCE, targets for the neutral class are manually extracted by annotators from each microblog, which are more closely related to the microblog content. Moreover, machine-translated test sets in both languages show worse F1-macro than the original test sets, indicating that machine translation fails to generate high-quality data. This suggests the importance of developing a zero-shot stance detection dataset for Chinese, which has not been done prior to this work.

[3] https://translate.google.com/

Train/Val	Test	XLNet	RoBERTa
M	N	.679	.688
M	C	.844	.846
C	N	.291	.254
N	C	.341	.342

Table 9: Comparison of $F1_{macro}$ of XLNet and RoBERTa using different types of targets for training/validation and test. M, N, and C represent data with mixed targets, noun-phrase targets, and claim targets, respectively.

Data	XLNet	RoBERTa
MB+C	.845	.838
C	.670	.678

Table 10: Comparison of $F1_{macro}$ of XLNet and RoBERTa when both microblog and claim target (MB+C) are used vs. when only claim target (C) is used as the input.

5.4 Impact of Incorporating Two Target Types

To analyze the impact of incorporating the noun-phrase targets and the claim targets in one dataset, we evaluate models trained with noun-phrase targets using the claim targets and vice versa. Results are compared with models trained using mixed target types and evaluated by two types of targets separately. Experiments are performed for subtask A, using the best-performing XLNet and RoBERTa.

Results are shown in Table 9, where we can observe that when models are trained with claim targets and evaluated with noun-phrase targets, the performance is much worse than ones trained by the mixed targets (e.g., 0.291 vs. 0.679 for XLNet). Similar results can be observed when models are trained with noun-phrase targets. These results suggest that datasets including the uni-target type are not capable of handling other target type, which further strengthens the necessity of developing datasets including both target types.

5.5 Spuriosity Analysis for Claim Targets

We perform spuriosity analysis for claim targets to ensure that we cannot predict the stance based

solely on the claim. For subtask A, we conduct experiments using XLNet and RoBERTa with only the claim target as input, which are compared with experiments using both microblog and claim target. The results are shown in Table 10, where we can observe a substantial amount of performance decrease when only the claim target is used as input. Therefore, both microblogs and claim targets are required for the models to make correct stance predictions by learning the semantic relation between them.

6 Conclusion

In this paper, we introduce C-STANCE, the first Chinese zero-shot stance detection dataset. Our dataset includes two challenging ZSSD subtasks: target-based ZSSD (evaluating classifiers using a large number of unseen targets) and domain-based ZSSD (evaluating classifiers using a large number of unseen targets from unseen domains). Moreover, we consider both noun-phrase targets and claim targets. Our dataset is larger and more challenging compared with the previous Chinese stance detection dataset, consisting of 48,126 annotated microblog-target pairs. C-STANCE can serve as a new benchmark for ZSSD, along with VAST, and can enable future research for other stance detection tasks. We conduct experiments using state-of-the-art deep learning models. Future work includes studying the multilingual ZSSD with the union of C-STANCE and other multi-lingual datasets.

Limitations

Our C-STANCE data is collected from social media, which may be seen as a limitation, as we may not cover all aspects of formal texts that could be used in essays or news comments. We will plan to extend this dataset with other types of text in the future. However, this is a limitation of any other datasets that focus on social media content.

Ethical Statement

Our dataset does not provide any personally identifiable information. Microblogs are collected using generic keywords instead of user information as queries, therefore our dataset does not have a large collection of microblogs from an individual user. Thus our dataset complies with Sina Weibo's information privacy policy.

Acknowledgements

We thank the National Science Foundation for support from grants IIS-1912887, IIS-2107487, and ITE-2137846 which supported the research and the computation in this study. We also thank our reviewers for their insightful feedback and comments.

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Inf. Process. Manage.*, 58(4).

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, Online. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020a. STANDER: An expert-annotated dataset for news stance detection and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4086–4101, Online. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020b. 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, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. hillary: What went viral during the 2016 us presidential election. In *Social Informatics*, pages 143–161, Cham. Springer International Publishing.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. 2015. Pheme: Computing veracity—the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, pages 3988–3994.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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, Online. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. 2020. Representativeness of abortion legislation debate on twitter: A case study in argentina and chile. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 765–774, New York, NY, USA. Association for Computing Machinery.

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)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

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, Hong Kong, China. Association for Computational Linguistics.

Yingjie Li and Cornelia Caragea. 2021a. A multi-task learning framework for multi-target stance detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2320–2326, Online. Association for Computational Linguistics.

Yingjie Li and Cornelia Caragea. 2021b. 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, Online. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021a. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021b. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1500–1509, New York, NY, USA. Association for Computing Machinery.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2738–2747, New York, NY, USA. Association for Computing Machinery.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

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*.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z. Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7112–7123, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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, Valencia, Spain. Association for Computational Linguistics.

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.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. An english-hindi code-mixed corpus: Stance annotation and baseline system. *arXiv preprint arXiv:1805.11868*.

Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. 2017. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. 2016. DeepStance at SemEval-2016 task 6: Detecting stance in tweets using character and word-level CNNs. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 413–419, San Diego,

California. Association for Computational Linguistics.

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*, SIGIR'19, page 1173–1176, New York, NY, USA. Association for Computing Machinery.

Penghui Wei, Wenji Mao, and Daniel Zeng. 2018. A target-guided neural memory model for stance detection in twitter. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California. Association for Computational Linguistics.

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)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural Language Understanding and Intelligent Applications*, pages 907–916, Cham. Springer International Publishing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 347–353, New York, NY, USA. Association for Computing Machinery.

A More Examples of C-STANCE

In this section, we show more examples for each domain of our C-STANCE dataset in Table 11.

B Query Keywords

The full keywords set that we used for data crawling is shown in Table 12. We generate the list by gradually extending the initial keywords set from *Weibo hot list* with the most frequent words.

C Token Overlap

We also report the average percentage of tokens in targets that overlap with tokens in microblogs. The results are shown in Table 16. We observe that noun-phrase show a higher overlapping percentage than claim targets, which is because annotators tend to summarize noun-phrase targets using semantically similar tokens from the text.

D Full Statistics of Subtask B

The statistics of the 7 dataset splits (data from six domains for training and validation, and the data from the left-out domain as the zero-shot test set) are shown in Table 13.

E Evaluations on Models Trained by Mixed Targets with Noun-Phrase Targets and Claim Targets

In subtask A and subtask B, for experiments using mixed targets, we also test the baseline models using noun-phrase targets and claim targets separately. Our goal is to better understand how each model trained on mixed targets performs for each type of target separately. The results for subtask A and subtask B are shown in Table 14 and Table 15, respectively. We can observe that the fine-tuned transformer-based models (i.e., BERT, RoBERTa, and XLNET) show higher performance on the claim targets. For BiCE, CrossNet, and TGA-Net, stances for claim targets are more difficult to predict.

CoE	Microblog	既然新冠要和人类长期共存，做疫苗接种应该比封控重要。再继续封锁，经济咋办呀？今天又要学校停课、企业停工。 Since the new crown will coexist with humans for a long time, vaccination should be more important than lockdown. If the blockade continues, what will happen to the economy? Today, schools are closed and businesses are closed.
	N target/Stance	封控 sealed management / Against
	C target/ Stance	新冠病毒会很快消失，不可能人类长期共存 The Covid-19 virus will disappear soon, and it is impossible for human beings to coexist for a long time. / Against
WE	Microblog	拜登表示如果普京愿意结束战争，他已做好了与普京对话的准备。美国搞残欧洲的目的达到，确实代理人战争再下去打没啥好处了。 Biden stated that if Putin is willing to end the war, he is ready to talk to Putin. The United States has achieved its goal of crippling Europe. It is true that there is no benefit in continuing the proxy war.
	N target/Stance	俄乌冲突 Russia-Ukraine conflict / Against
	C target/ Stance	美国搞乱欧洲的目的达到了，俄乌战争没有什么好打的了。 The purpose of the United States to mess up Europe has been achieved, and it's meaningless to continue fighting the Russia-Ukraine war. / Favor
CuE	Microblog	我知道公立教育的问题，但我还是感谢公立教育让我和来自比我有钱家庭的人在一起享受同样教育。 I know the problems with public education, but I'm still thankful that public education allows me to enjoy the same education with people from families richer than me.
	N target/Stance	公立教育 public education / Favor
	C target/ Stance	真正有钱的家庭都在私立学校，教育公平的天平秤早就倾斜了。 Children from truly wealthy families are all in private schools, and the balance of educational equity has long been tilted. / Against
EC	Microblog	发现沉沦短视频时代久了我无法静心去读一段文字，逃避阅读长文。阅读也走马观花。 I found that after watching too many short videos, I can't read a paragraph of text quietly and avoid reading long texts. Reading for me is a quick glance.
	N target/Stance	短视频 short video / Against
	C target/ Stance	短视频做得最好的应该就是抖音了。 TicTok is the best short video platform. / Neutral
S	Microblog	昨晚云达女足的比赛来到威悉球场进行，20417名梅粉来到现场！虽然1比2惜败，但我们听到了你们的声音！ Last night, the Werder women's football game came to the Weser Stadium, and 20,417 fans came to the scene! Although they lost the game by 1-2, we heard your voices!
	N target/Stance	运达女足 Werder Women's Football / Favor
	C target/ Stance	这么多球迷去看运达女足的比赛，结果输了，也太让球迷们失望了吧。 So many fans went to watch the Yunda women's game, but they lost. So disappointing for the fans. / Against
R	Microblog	体质占优势的男性就掌握了话语权。所以实现真正的女权发展科技，当科技可以抹平和男性的生产力和武力差距后，才能真正实现男女平等。 Men with dominant physiques have the right to speak. Therefore, to realize true women's rights and develop technology, only when technology can bridge the gap in productivity and force with men can we truly achieve equality between men and women.
	N target/Stance	男女平等 gender equality / Favor
	C target/ Stance	女权只需要嘴巴说说就好了，无需行动，时间可以改变一切。 Women's rights only need to be talked about, no action needed, time can change everything. / Against
EP	Microblog	延缓气候变化，需要富裕国家更多采取更多行动。在澳洲，还是有很多人对气候变暖持怀疑态度，这也阻碍了政府采取更多行动。 Slowing climate change will require rich countries to do more. In Australia, there are still many people who are skeptical about climate change, which is also preventing the government from taking more action.
	N target/Stance	延缓气候变化 Slow down climate change / Favor
	C target/ Stance	气候变化是二氧化碳等气体变多导致的，其造成的后果也很大 Climate change is caused by the increase of gases such as carbon dioxide, and its consequences are also large. / Neutral

Table 11: Examples of noun-phrase targets and claim targets for microblogs in each domain of our C-STANCE dataset. "N target" and "C target" represent the noun-phrase target and the claim target, respectively.

股市 stock market, 读书 read, 艺术 art, 设计 design, 男朋友 boyfriend, 文化输出 cultural output, 社会 society, 父母 parents, 消费观 consumption concept, 战争 war, 异地恋 long distance relationship, 带娃 raise a baby, 女朋友 girlfriend, 大学 college, 华语乐坛 Chinese pop music, 健身 fitness, 电影 movie, 播客 podcast, 公立教育 public education, 世界新闻 world news, 加强针 booster, 疫苗 vaccine, 气候变化 climate change, 人工智能 artificial intelligence, 书 book, LGBTQ,拆迁 remove, 俄罗斯 Russia, 乌克兰 Ukraine, 汽油 gasoline, 武器 wearpons, 大选 general election, 知识 knowledge, 选举 election, 口罩 face mask, 鸡娃 force kids to compete, 高中 high school, 贫困 poverty, 财经 financial, 育儿 parenting, 规划人生 life planning, 男篮 men's basketball,高校 colleges and universities, 男足 men's football, 女足 women's football, 教师 teacher, 思考 thinking, 医患 doctors and patients, 军事 military, 人口负增长 negative population growth, 篮球 basketball, 辩论 debate, 环境 environment, 总统 president, 学生 student, 婚姻 marriage, 科学 science, 医保 medical insurance,封控 sealed management, 保险 insurance, 工作 work, 油价 oil price, 防疫 epidemic prevention, 世界杯 World Fup, NBA, 男女平等 gender equality, 平权 equal rights, 移民 migrant, 新冠疫苗 Covid-19 vaccine, 直播带货 livestream shopping, 短视频 short video, 物价 prices, 流行文化 pop culture, 自由恋爱 free love, 相亲 blind date, 素质教育 quality education, 中医 traditional Chinese medicine, 静默 silence, 新冠共存 co-existence with coronavirus, 上网课 online class, 居家办公 work from home, 电商 e-commerce, 女拳 women's rights, iphone, 新能源 new energy, 垃圾分类 garbage classification, 微商 Wechat business, 中国防疫 China's epidemic prevention, 防控 prevention and control, 老龄化 population aging, 中国历史 Chinese history, 传统文化 traditional culture, 近代史 modern history, 阅读 read, 芯片 chip, 投资 invest, 电视剧 TV series, 影评 movie review, 票房 box office,高考 college entrance examination 美妆博主 beauty blogger, 足球 football, 体育 sports, 健康 healthy, 群体免疫 herd immunity, 减负 lighten the burden, 农村 the countryside, 环保意识 environmental awareness

Table 12: The full query keywords list used in our work for microblog crawling.

		# Examples		# Targets		Avg. Length			# Unique MB
		N	C	N	C	N	C	MB	
Covid Epidemic	Train	12,379	18,984	7,519	18,585	4.0	26.0	102.4	6,690
	Val	2,249	3,447	2,208	3,436	4.6	26.0	104.8	1,167
	Test	3,474	5,346	1,896	5,211	3.7	25.7	103.6	1,786
World Event	Train	11,978	18,417	7,426	18,034	4.0	25.9	101.6	6,813
	Val	2,077	3,186	2,045	3,176	4.6	26.0	104.7	1,087
	Test	3,130	4,770	2,152	4,673	4.2	26.1	105.8	1,591
Culture and Education	Train	12,283	18,720	7,671	18,314	4.0	26.0	102.8	7,105
	Val	2,180	3,354	2,146	3,342	4.6	26.0	104.7	1,131
	Test	2,397	3,618	1,806	3,589	3.9	25.6	104.0	1,218
Entertainment and consumption	Train	10,517	16,110	6,777	15,811	4.1	26.1	103.6	6,244
	Val	1,991	3,051	1,960	3,042	4.7	26.0	106.6	1,043
	Test	4,010	6,153	2,886	6,042	3.9	25.6	98.9	2,052
Sports	Train	13,549	20,682	8,091	20,237	3.9	25.9	103.4	7,379
	Val	2,321	3,558	2,276	3,548	4.6	26.0	105.2	1,192
	Test	2,088	3,177	1,256	3,117	3.8	25.7	96.8	1,060
Rights	Train	12,797	19,548	7,793	19,146	4.0	25.8	102.2	7,094
	Val	2,352	3,594	2,307	3,583	4.6	26.0	104.6	1,218
	Test	2,492	3,828	1,523	3,728	3.8	26.6	105.4	1,276
Environmental Protection	Train	14,237	21,882	8,246	21,404	3.9	25.9	102.1	7,708
	Val	2,363	3,636	2,321	3,626	4.6	25.8	104.4	1,223
	Test	1,453	2,196	1,056	2,131	4.4	26.9	107.7	733

Table 13: Data statistics of all 7 dataset splits for subtask B. N, C, and MB represent noun-phrase targets, claim targets, and microblogs, respectively.

	Mixed targets				Noun-phrase targets				Claim targets			
	Con	Pro	Neu	All	Con	Pro	Neu	All	Con	Pro	Neu	All
BiCE	.490	.408	.443	.447	.518	.544	.562	.541	.476	.302	.354	.377
Cross-Net	.526	.541	.592	.553	.582	.571	.551	.568	.487	.52	.616	.541
TGA Net	.565	.599	.637	.600	.644	.629	.586	.620	.518	.577	.666	.587
BERT	.758	.763	.798	.773	.686	.679	.628	.665	.800	.828	.896	.841
RoBERTa	**.775**	**.769**	**.811**	**.785**	.712	**.692**	**.659**	**.688**	**.813**	.826	.899	**.846**
XLNet	.767	**.769**	.804	.780	**.715**	.683	.640	.679	.800	**.831**	**.902**	.844

Table 14: Comparison of different models in subtask A, which are trained on mixed targets and tested using the full test set with mixed targets (M), the noun-phrase targets (N), and the claim targets (C), respectively. Results are averaged over four runs.

Model		CoE	WE	CuE	EC	S	R	EP
BiCE	M	.347	.413	.376	.393	.413	.360	.400
	N	.428	.551	.478	.509	.545	.464	.509
	C	.296	.322	.310	.319	.327	.295	.328
CrossNet	M	.374	.375	.370	.392	.374	.351	.386
	N	.484	.561	.490	.503	.534	.475	.527
	C	.263	.217	.253	.279	.212	.229	.245
TGA-Net	M	.570	.581	.598	.598	.609	.608	.592
	N	.545	.586	.603	.599	.618	.602	.585
	C	.576	.571	.585	.589	.607	.587	.595
BERT	M	.753	.773	.768	.762	.775	.772	.777
	N	.598	**.665**	.629	.645	.677	.627	.619
	C	.829	.836	.839	.832	.838	.836	.873
RoBERTa	M	.755	**.776**	**.779**	**.774**	**.785**	**.784**	**.795**
	N	.596	.655	**.650**	**.668**	**.683**	**.647**	**.657**
	C	**.834**	**.848**	**.845**	.835	**.850**	**.847**	**.879**
XLNet	M	**.758**	.763	.778	.767	.777	.777	.781
	N	**.604**	.649	**.650**	.650	.677	**.647**	.642
	C	.829	.833	**.845**	**.836**	.841	.831	.870

Table 15: Comparison of $F1_{macro}$ of different models trained on mixed targets for 7 different zero-shot domain settings, and tested using the full test set with mixed targets (M), the noun-phrase targets (N), and the claim targets (C), respectively. Results are averaged over four runs.

	N	C
Train	78.2%	25.5%
Val	76.4%	24.7%
Test	77.6%	25.8%

Table 16: Average percentage of token overlap between two types of targets and microblogs. N and C represent noun-phrase targets and claim targets, respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation Section after the Conclusion.

☑ A2. Did you discuss any potential risks of your work?
Limitation Section after the Conclusion.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 and 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 3 and 4.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We discuss the dataset that we created in Section 3 and the baseline models that we used in Section 4.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We discussed our usage of baseline models in Section 4. We discussed the intended use of the dataset that we created in Section 3. We show our dataset is compatible with the original access conditions in the Ethical Statement Section.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethical Statement Section.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.

C ☑ Did you run computational experiments?

Section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We discussed the annotation company that we worked with and how we recruited annotators in Section 3.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3 and Section 5.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.