EZ-STANCE: A Large Dataset for English Zero-Shot Stance Detection

Chenye Zhao **Cornelia Caragea**
Computer Science
University of Illinois Chicago
czhao43@uic.edu cornelia@uic.edu

Abstract

Zero-shot stance detection (ZSSD) aims to determine whether the author of a text is in favor, against, or neutral toward a target that is unseen during training. In this paper, we present EZ-STANCE, a large **E**nglish **Z**SSD dataset with 47,316 annotated text-target pairs. In contrast to VAST (Allaway and McKeown, 2020), which is *the only other large existing ZSSD dataset for English*, EZ-STANCE is 2.5 times larger, includes both noun-phrase targets and claim targets that cover a wide range of domains, provides two challenging subtasks for ZSSD: target-based ZSSD and domain-based ZSSD, and contains much harder examples for the *neutral* class. We evaluate EZ-STANCE using state-of-the-art deep learning models. Furthermore, we propose to transform ZSSD into the NLI task by applying simple yet effective prompts to noun-phrase targets. Our experimental results show that EZ-STANCE is a challenging new benchmark, which provides significant research opportunities on English ZSSD. We publicly release our dataset and code at `https://github.com/chenyez/EZ-STANCE`.

1 Introduction

The goal of stance detection is to automatically detect 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; AL-Dayel and Magdy, 2021), e.g., public education, mask mandate, or nuclear energy. The stance can reveal valuable insights relevant to events such as public policy-making and presidential elections.

Earlier research has concentrated on two types of stance detection tasks: in-target stance detection, in which models are trained and evaluated using data from the same target, e.g., both train and test sets contain data about "Donald Trump" (Hasan and Ng, 2014; Mohammad et al., 2016b; Graells-Garrido et al., 2020), and cross-target stance detection, where the models are trained on source targets that are related to, but distinct from, the destination targets (Augenstein et al., 2016; Wei and Mao, 2019), which remain unseen during training (e.g., the destination target is "Donald Trump" whereas the source target is "Hillary Clinton"). However, it is unrealistic to incorporate every potential or related target in the training set. As such, zero-shot stance detection (ZSSD) has emerged as a promising direction (Allaway and McKeown, 2020) to evaluate classifiers on a large number of unseen (and unrelated) targets. ZSSD is more related to real-world scenarios and has consequently started to receive significant interest recently (Liu et al., 2021; Luo et al., 2022; Liang et al., 2022b).

Despite the growing interest in ZSSD, the task still exhibits several limitations. First, the VAST dataset (Allaway and McKeown, 2020) which is the only other large existing ZSSD dataset for English, contains only noun phrase targets. Yet, in real-world scenarios, stance is often taken toward both noun phrases (Mohammad et al., 2016b; Glandt et al., 2021) and claims (Ferreira and Vlachos, 2016; Derczynski et al., 2017). We notice that models trained on data with noun-phrase targets struggle to accurately predict the stance for claim-target data, and vice versa, due to the mismatch between the training and test data. The need to incorporate both types of targets for ZSSD has been relatively overlooked. Second, VAST is designed solely to detect the stance of unseen targets, but these unseen targets at the inference stage originate from the same domain as the training targets (in-domain), possessing similar semantics, which makes the task less challenging. Third, despite being instrumental for the development of zero-shot stance detection, VAST generates data for the neutral class by randomly permuting documents and targets, leading to a lack of semantic correlation between the two (we show an example from VAST from the neutral class in Table 1). Deep learning models can easily detect these patterns, consequently diminishing the

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15697–15714
August 11-16, 2024 ©2024 Association for Computational Linguistics

VAST

Text: So if someone can't do algebra they can out of it, but if another student takes it and fails they get an F on their transcripts? How could this work and be fair to those who attempt to take subjects which challenge themselves?
Stance/Noun-phrase targets: Neutral / medical website

EZ-STANCE

Text: What happened to "herd immunity"? Are people supposed to hide under their beds in a zip lock bag?
Stance/Noun-phrase targets: Against / herd immunity
Stance/Claim targets: Favor / People are not supposed to be forced to stay indoors.
Stance/Noun-phrase targets: Neutral / zip lock bag

Table 1: Examples from EZ-STANCE and VAST.

complexity of the task.

In an effort to address the aforementioned limitations and spur research in ZSSD, we present EZ-STANCE, a large **E**nglish **Z**ero-shot stance detection dataset collected from Twitter. In contrast with VAST, EZ-STANCE is, to our knowledge, the first large English ZSSD dataset that captures both noun-phrase targets and claim targets, covering a more diverse set of targets. By training a single model on our comprehensive dataset, we achieve comparable or superior performance than training separate models for each type of target. Moreover, EZ-STANCE includes two real-world scenarios for zero-shot stance detection, namely target-based and domain-based ZSSD. **Subtask A: target-based zero-shot stance detection**. This subtask is the same as the traditional ZSSD task (Allaway and McKeown, 2020), where stance detection classifiers are evaluated using a large number of completely unseen (and unrelated) targets, but from the same domains (in-domain). **Subtask B: domain-based zero-shot stance detection**. This subtask is our proposed ZSSD task where stance detection classifiers are evaluated using a large number of unseen targets from completely new domains (out-of-domain). Furthermore, in EZ-STANCE, annotators manually extract targets from each tweet to form the neutral class, ensuring semantic relevance to the tweet content. We show an example from our dataset along with corresponding noun-phrase and claim targets for each stance (against, favor, and neutral) in Table 1. As we can see from the table, the noun-phrase target "zip lock bag" is relevant to the tweet but the author of the tweet holds a neutral stance towards this target.

In summary, our contributions are as follows: 1) We present EZ-STANCE, a unique large zero-shot stance detection dataset, composed of 47,316 annotated English tweet-target pairs. EZ-STANCE is 2.5 times larger than VAST (Allaway and McKeown, 2020), which is the only large existing ZSSD dataset for English. We provide a detailed description and analysis of our dataset; 2) We consider a more diverse set of targets including both noun phrase and claim targets (see Table 1); 3) We include two challenging ZSSD subtasks in EZ-STANCE: target-based zero-shot stance detection and domain-based zero-shot stance detection; 4) We establish baseline results using both traditional models and pre-trained language models; 5) We propose to formulate stance detection into the task of natural language inference (NLI) by applying simple yet effective prompts to noun-phrase targets. Our results and analysis show that EZ-STANCE is a challenging new benchmark.

2 Related Work

Target-specific stance detection is the most prevalent type of stance detection (ALDayel and Magdy, 2021), whose goal is to determine the stance expressed in a text towards a target. Usually, the target is *an entity / short noun-phrase*, e.g., a political figure or controversial topic (Hasan and Ng, 2014; Mohammad et al., 2016a; Zotova et al., 2020; Conforti et al., 2020a,b), or *a claim*, e.g., an article's headline or a reply to a rumorous 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).

Most earlier research is centered around 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, 2021). However, the challenge often arises in gathering enough annotated data for each specific target, and traditional models perform poorly when generalized to unseen target data. This spurred interest in investigating cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Wei and Mao, 2019; Zhang et al., 2020), where a classifier is adapted from different but related targets. However, cross-target stance detection still requires 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 (Liang et al., 2022b).

Zero-shot stance detection (ZSSD) which aims

Name	Authors	Source	# Target(s)	Target Type		Task	Size
				NounPhrase	Claim		
SemEval16	Mohammad et al. (2016b)	Twitter	6	✓	✗	In-target	4,870
MultiTarget	Sobhani et al. (2017)	Twitter	4	✓	✗	Cross-target	4,455
WTWT	Conforti et al. (2020b)	Twitter	5	✓	✗	In-target	51,284
VAST	Allaway and McKeown (2020)	Comments	5,634	✓	✗	ZSSD	18,545
Covid19	Glandt et al. (2021)	Twitter	4	✓	✗	In-target	6,133
P-STANCE	Li et al. (2021)	Twitter	3	✓	✗	In-target	21,574
Emergent	Ferreira and Vlachos (2016)	News articles	300	✗	✓	ZSSD	2,595
RumEval17	Derczynski et al. (2017)	Twitter	325	✗	✓	Cross-target	5,568
RumEval19	Gorrell et al. (2019)	Twitter, Reddit	446	✗	✓	ZSSD	8,574
EZ-STANCE	**Ours**	Twitter	40,678	✓	✓	ZSSD	47,316

Table 2: Comparison of English stance detection datasets.

to detect the stance on a large number of unseen (and unrelated) targets represents a more realistic scenario, and has received significant interest in recent years (Allaway and McKeown, 2020; Liu et al., 2021; Liang et al., 2022a,b; Luo et al., 2022; Xu et al., 2022; Li et al., 2023; Zhao et al., 2023). These works tackle stance detection where the target is an entity or noun phrase, whereas other works that follow a ZSSD scenario (training and test data are collected from different events) focus on stance classification towards rumours in which the target is a longer text or a claim (Ferreira and Vlachos, 2016; Derczynski et al., 2017; Gorrell et al., 2019).

Surprisingly, despite substantial interest in ZSSD, large-scale datasets for the evaluation of this task are limited. In Table 2, we compare our EZ-STANCE dataset with existing English stance detection datasets. As we can observe from the table, VAried Stance Topics (VAST) (Allaway and McKeown, 2020) is the only existing dataset for English ZSSD that encompasses thousands of targets, whereas Emergent (Ferreira and Vlachos, 2016), RumEval17 (Derczynski et al., 2017), RumEval19 (Gorrell et al., 2019) contain only several hundred targets. We can also observe that these datasets have either noun phrase targets or claim targets, but not both. In contrast, our EZ-STANCE dataset integrates *both target types* into a single dataset; includes two types of ZSSD subtasks—*target-based ZSSD* (same as VAST) and *domain-based ZSSD*, a new and more challenging task in which classifiers are evaluated on unseen targets from completely new domains; contains much harder *neutral* examples—the targets of the neutral data in EZ-STANCE are extracted from the texts, ensuring strong semantic relevance to the text content, while data for the neutral class in VAST is generated by randomly permuting existing texts and targets, resulting in easy-to-detect patterns; and is much larger in size—EZ-STANCE is notable for its extensive range of 40,678 targets across a comprehensive corpus of 47,316 examples (we provide more examples from EZ-STANCE in Appendix A).

3 Dataset Construction

Here, we detail the construction of EZ-STANCE.

3.1 Data Collection

Our data are collected using the Twitter API, spanning from May 2021 to January 2023. Like previous works (Mohammad et al., 2016b; Glandt et al., 2021; Li et al., 2021), we crawl tweets using query keywords. To cover a wide range of domains, we use domain names from the *Explore* page of Twitter as keywords for crawling (e.g., sports, education, etc.). Then we expand the keywords set for the next round by including the most frequent words as supplementary keywords. In total, we collect 50,000 tweets. Next, we perform data filtering to eliminate keywords and tweets that are not suitable for stance detection. The detailed keywords selection process, the full list of keywords for crawling, as well as our data filtering strategy are provided in Appendix B. Eventually, we select 72 keywords covering controversial topics. We summarize the 72 keywords into 8 domains: “Covid Epidemic" (CE), “World Events" (WE), “Education and Culture" (EdC), “Entertainment and Consumption" (EnC), “Sports" (S), “Rights" (R), “Environmental Protection" (EP), and “Politics" (P). Table 3 shows the domains and query keywords in each domain.

3.2 Data Annotation

The target and stance annotations of our dataset are gathered through Cogitotech,[1] a data annotation company that provides annotation services for big AI companies (e.g., OpenAI, AWS, etc.). To

[1] https://www.cogitotech.com/

Domain		Query Keywords
Covid Epidemic	CE	epidemic prevention, living with covid, herd-immunity, WFH, booster, vaccine, mask mandate, FDA, post-covid, Fauci
World Events	WE	world news, Ukraine, Russia, migrant, NATO, China, Mideast, negative population growth, terrorism
Education and Culture	EdC	public education, pop culture, cultural output, home schooling, AI assistance writing, arming teachers, private education, international student
Entertainment and Consumption	EnC	prices, gasoline price, online shopping, TikTok, iPhone, Reels, Disney, medical insurance, ethical consumption, vegetarian
Sports	S	World Cup, NBA, men's football, women's football, NCAA, MLB, NFL, WWE
Rights	R	gender equality, equal rights, women's rights, LGBTQ, BLM, doctors and patients, racism, Asian hate, gun control
Environmental Protection	EP	climate change, clean energy, environmental awareness, environmental protection agency, shut down coal plants, nuclear energy, electric vehicle
Politics	P	government, republican, reform, leftists, democrat, democracy, right-wing, politic, presidential debate, presidential election, midterm election

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

	Noun-phrase targets			Claim targets		
Domain	Con	Pro	Neu	Con	Pro	Neu
CE	971	812	853	1,329	1,328	1,327
WE	856	559	850	1,140	1,139	1,140
EdC	615	826	647	1,083	1,083	1,083
EnC	636	925	1,084	1,405	1,406	1,405
S	179	781	808	941	942	941
R	910	1,015	522	1,191	1,192	1,191
EP	515	987	563	979	980	979
P	1,184	846	829	1,386	1,387	1,386
Overall	5,866	6,751	6,156	9,454	9,457	9,452

Table 4: Overall label distribution for noun-phrase and claim targets in each domain from our dataset. Con, Pro, Neu represent against, favor, and neutral, respectively.

ensure high-quality annotations, we apply rigorous criteria: 1) Annotators must have a minimum education qualification of college graduation; 2) The annotators' native language must be English. Moreover, we randomly sample 10% of each annotator's annotations to perform quality checks and discard annotations from an annotator if the acceptance rate is lower than 90%. This data is re-sent to other qualified annotators for labeling. The overall stance label distribution for both noun-phrase and claim targets for each domain is shown in Table 4.

3.2.1 Annotation for Noun-Phrase Targets

The annotation for noun-phrase targets is performed in two steps. In step 1, one annotator is asked to identify a minimum of 2 targets from each given tweet. In step 2, we instruct 3 annotators to assign a stance label to each tweet-target pair. The instructions for the annotators are provided in Appendix C.1. After the annotations are completed, we determine the stance for each tweet-target pair by using the majority vote amongst the three annotators. The inter-annotator agreement measured using Krippendorff's alpha (Krippendorff, 2011) is 0.63, which is higher than VAST (0.427).

3.2.2 Annotation for Claim Targets

We ask one annotator to generate three claims for each tweet, to which the tweet takes favor, against, and neutral stances, respectively. The detailed annotation instructions are in Appendix C.2. For quality assurance, we hide the stance labels for a subset of tweet-claim pairs and ask another group of annotators (who did not write the claims) to annotate the stance. The two groups agree on 95% of the times. This result indicates high-quality generations of the claim targets and stance labels.

3.3 Dataset Split

For subtask A, we split the dataset in alignment with the VAST dataset (Allaway and McKeown, 2020): the training, validation, and test sets do not share any texts (tweets) and targets with each other. The detailed split process are shown in Appendix D. For subtask B, we use the data from seven domains (source) for training and validation, and the data from the left-out domain (zero-shot) as the test set. This results in 8 dataset splits for subtask B with one dataset split assigned for each of the eight domains, wherein each domain in turn is used as the test set. We exclude data with overlapping targets between the source and zero-shot domains, and then partition the source domains into training and validation sets, ensuring no duplication of tweets and targets. The statistics of subtask A and subtask B (using "Covid Epidemic" as the zero-shot domain) are shown in Table 5. The full statistics of subtask B are shown in Appendix E.

3.4 Dataset Statistics

In this section, we present a statistical analysis of our EZ-STANCE dataset.

		# Examples		# Unique			Avg. Length			Lexsim
		N	C	N	C	T	N	C	T	(%)
EZ-STANCE Subtask A	Train	13,756	18,879	7,437	18,861	6,293	1.8	19.0	40.0	-
	Val	2,354	4,349	2,284	4,345	1,454	2.4	19.0	40.0	13
	Test	2,663	5,135	2,621	5,130	1,715	2.4	19.3	39.7	12
EZ-STANCE Subtask B (CE)	Train	12,648	19,467	8,506	19,440	6,489	2.0	18.9	39.4	-
	Val	1,958	3,753	1,932	3,749	1,251	2.4	19.2	40.4	11
	Test	2,639	3,819	1,734	3,814	1,273	1.9	19.2	41.8	10
VAST	Train	13,477	-	4,003	-	1,845	1.8	-	101.3	-
	Val	1,019	-	383	-	682	2.3	-	99.8	19
	Test	1,460	-	600	-	786	2.4	-	100.5	16

Table 5: Comparison of key statistics of EZ-STANCE subtask A and subtask B (with Covid Epidemic (CE) as the zero-shot domain) with the existing English ZSSD dataset (VAST). N, C, T represent noun-phrase targets, claim targets, and texts/tweets, respectively. Lexsim represents the percentage of LexsimTopics.

Label Distribution We can observe from Table 4 that the "Sports" (S) and the "Entertainment and Consumption" (EnC) domains have the highest percentage in the "Neutral" class. This might be because these domains include more tweets related to news. Moreover, people are showing a higher percentage of "Against" stances toward targets in the "Covid Epidemic" (CoE), "World Event" (WE), and "Politics" (P) domains, where more contrary opinions are often expressed.

Dataset Size In Table 5, we observe that EZ-STANCE includes a much larger number of zero-shot targets than VAST. EZ-STANCE uniquely provides zero-shot claim targets, further expanding its coverage. Moreover, VAST only includes target-based ZSSD, whereas EZ-STANCE also enables the more challenging domain-based ZSSD.

Text/target Lengths Table 5 indicates that the average word counts in EZ-STANCE are 2 for noun-phrase targets, 19 for claim targets, and 40 for texts. VAST features similarly lengthed noun-phrase targets but longer texts (around 100 words), which are from New York Times news comments, unlike the shorter tweet-based EZ-STANCE texts.

LexSimTopics Given the linguistic variations in the noun-phrase target expressions, we investigate the prevalence of *LexSimTopics* (Allaway and McKeown, 2020) between the train and test sets. LexSimTopics is defined as the percentage of targets that possess more than 0.9 cosine similarities with any training targets in the word embedding space (Bojanowski et al., 2017). As shown in Table 5, in Subtask A, we have 12% and 13% *LexSimTopics* in the test set and the validation set, respectively, whereas for the "Covid Epidemic" domain in subtask B, we only have 10% and 11% *LexSimTopics* for the test and validation sets, indicating that subtask B poses more challenges as the targets in the training and test sets exhibit more differences. In comparison, the VAST dataset has 16% and 19% *LexSimTopics* in the test set and validation set, respectively, which are higher than our dataset.

4 Methodology

We now present our approach for converting ZSSD into the natural language inference (NLI) task.

4.1 Problem Definition

Suppose we are given a training set D^{train}= $\{(x_i^{train}, t_i^{train}, y_i^{train})\}_{i=1}^{N_{train}}$ and a test set D^{test}= $\{(x_i^{test}, t_i^{test})\}_{i=1}^{N_{test}}$, where x_i^{train} is a training document (tweet), t_i^{train} is a target and y_i^{train} is the label (or stance) $\in \{Favor, Against, Neutral\}$. For target-based ZSSD (subtask A), targets in x_i^{test} do not overlap with targets in x_i^{train}. For domain-based ZSSD (subtask B), targets in x_i^{test} not only do not overlap with the targets in x_i^{train}, but they also belong to a domain that is not seen in D^{train}. The objective is to predict the stance given both x_i^{test} and t_i^{test} by training a model on the D^{train}.

4.2 Transform ZSSD into NLI

Natural Language Inference (NLI) (Bowman et al., 2015; Williams et al., 2018) is a task that classifies the relationship between a premise and a hypothesis as entailment, contradiction, or neutral. Models pre-trained on NLI datasets are adept at discerning intricate relationships between sentences, a skill that is beneficial for similar NLP tasks.

To leverage the extensive knowledge of NLI pre-trained models, we propose to transform stance detection into NLI. Particularly, we convert the text and target into the premise and hypothesis, respectively. The task of predicting stance labels (*Favor*, *Against*, or *Neutral*) is transformed into the task of predicting entailment labels (*Entailment*, *Contradiction*, or *Neutral*). To effectively apply

Texts (Premise):	Nuclear Energy is a much safer and cost efficient source of energy than coal and oil and people should be using it.
Target:	Nuclear Energy
Stance:	Favor
Hypothesis:	The premise has an entailment relation with Nuclear Energy!
NLI Label:	Entailment
Texts (Premise):	after struggling with my medical insurance for months i finally got an appointment win an endocrinologist to start.
Target:	medical insurance
Stance:	Against
Hypothesis:	The above text entails medical insurance!
NLI Label:	Contradiction

Table 6: Examples of formulating ZSSD as NLI. The prompts for NLI hypothesis formulation are in red.

NLI pre-trained models for stance detection, we introduce prompt templates that transform noun-phrase targets into sentence-like hypotheses, aligning them with the typical sentence format of NLI hypotheses. Claim targets remain the same as they already resemble hypotheses quite closely. Particularly, we design five simple yet effective prompts: "*The above text entails [target]!*", "*The premise has an entailment relation with [target]!*", "*This implies an entailment relation with [target]!*", "*The premise has the entailment relation with the hypothesis [target]!*", and "*The premise entails the hypothesis [target]!*". For each noun-phrase target, we randomly apply one of the five prompts. Examples in Table 6 demonstrate the re-formulation of ZSSD as NLI. We fine-tune the BART-large encoder (Lewis et al., 2020) pre-trained on MNLI (Williams et al., 2018) dataset to predict the stance.

5 Baselines and Models

We introduce the ZSSD baselines and our approach of utilizing NLI pre-trained models for ZSSD.

5.1 ZSSD Baselines

We evaluate EZ-STANCE using the following ZSSD baselines. *BiCE* (Augenstein et al., 2016) and *CrossNet* (Xu et al., 2018) predict the stance using the conditional encoding of BiLSTM. *TGA-Net* (Allaway and McKeown, 2020) captures implicit relations/correlations between targets in a hidden space to assist stance classification. Next, we consider fine-tuning the base version of state-of-the-art transformer-based models as strong baselines, including *BERT* (Devlin et al., 2019), *RoBERTa* (Liu et al., 2019) and *XLNet* (Yang et al., 2019). We also use *LLaMA 2* and *ChatGPT* to directly predict the stance label based on a task description (Zhang et al., 2023; Touvron et al., 2023).

5.2 NLI Pre-trained Models

To evaluate NLI pre-trained models for ZSSD, we compare the following methods. *BART-MNLI-e*: We fine-tune the BART-MNLI encoder on our original EZ-STANCE dataset *without prompts*. The BART decoder is not included due to memory constraints. *BART-MNLI-e$_p$*: Same as above, except that the BART-MNLI encoder is fine-tuned on EZ-STANCE *with prompts* applied to noun-phrase targets. Note that *BART-MNLI-e* and *BART-MNLI-e$_p$* are identical in experiments conducted with claim targets. *BART-MNLI*: We use the pre-trained BART-MNLI online version with both encoder and decoder *without fine-tuning on EZ-STANCE* to infer the stance labels for the test set of EZ-STANCE. In addition, to verify our approach across different model architectures, we fine-tune other NLI pre-trained models including *BERT-MNLI*, *RoBERTa-MNLI*, and *XLNet-MNLI* using the original EZ-STANCE dataset. We also experiment with these models using prompted noun-phrase targets, designated as *BERT-MNLI$_p$*, *RoBERTa-MNLI$_p$*, and *XLNet-MNLI$_p$*, respectively. We show the hyperparameters used in experiments in Appendix F.

6 Results

In this section, we first present results for subtask A (§6.1) and subtask B (§6.2). We then compare EZ-STANCE with the VAST dataset (§6.3). Next, we study the impact of different prompt designs (§6.4). Like prior works (Allaway and McKeown, 2020), we employ the macro-averaged F1 score across all classes as our evaluation metrics.

6.1 Target-based Zero-Shot Stance Detection

Target-based ZSSD (subtask A) aims to evaluate the classifier on a large number of completely unseen targets. We train models using three scenarios: 1) on the full training set with both noun-phrase and claim targets; 2) on training data with noun-phrase targets only; and 3) training data with claim targets only. Each model is then evaluated in three corresponding scenarios: 1) the full test set with mixed targets; 2) the test subset with noun-phrase targets only; and 3) the test subset with claim targets only.

Results are shown in Table 7. First, we observe that models trained on noun-phrase targets and evaluated on claim targets (N→C), or the re-

Train/Val	Mixed targets (M)			Noun-phrase targets (N)			Claim targets (C)		
Test	M	N	C	M	N	C	M	N	C
BiCE	.468	.476	.451	.398	.529	.334	.286	.282	.316
Cross-Net	.518	.509	.522	.407	.551	.336	.447	.241	.523
TGA Net	.590	.579	.594	.398	.606	.334	.286	.282	.596
LLaMA 2	.404	.436	.374	.404	.436	.374	.404	.436	.374
ChatGPT	.499	.604	.440	.499	.604	.440	.499	**.604**	.440
RoBERTa	.784	.639	.858	.433	.656	.318	.687	.345	.856
RoBERTa-MNLI	.797	.642	.876	.525	.659	.451	.706	.339	.880
RoBERTa-MNLI$_p$	.799	.661	.878	.532	.662	.492	-	-	-
BART-MNLI	.664	.295	.817	**.664**	.295	**.817**	.664	.295	.817
BART-MNLI-e	.810	.661	.883	.451	.675	.334	**.716***	.345	**.888***
BART-MNLI-e$_p$	**.812***	**.669***	**.885***	.446	**.687***	.322	-	-	-

Table 7: Subtask A: Comparison of F1$_{macro}$ of models on EZ-STANCE. $*$: our approach improves the best ZSSD baseline at $p < 0.05$ with paired t-test.

verse (C→N), demonstrate much worse performance compared to those trained with mixed targets (M→N or M→C) or with consistent target types (N→N or C→C). This suggests that datasets limited to one target type struggle to correctly predict the stance of the other, highlighting the necessity for developing a dataset that includes both noun-phrase and claim targets. Second, models trained on mixed targets demonstrate similar performance on both noun-phrase and claim targets (M→N or M→C) compared to those trained solely on noun-phrase or claim targets (N→N or C→C), respectively. This underscores the advantages of our dataset: training a single model on mixed targets yields similar results compared to training separate models for each target type, demonstrating both efficiency and efficacy. Third, fine-tuning MNLI pre-trained models (e.g., BART-MNLI-e, etc.) consistently outperform traditional ZSSD baselines (that do not have NLI pre-trained knowledge) in all three settings, showing the effectiveness of transforming ZSSD into NLI. Forth, when evaluated on noun-phrase targets, MNLI models trained with prompted noun-phrase targets consistently outperform those trained with original noun-phrase targets, indicating that our proposed prompts can effectively formulate noun-phrase targets into more refined hypotheses to better leverage the MNLI model for ZSSD. Note that when trained with claim targets, where prompts are not incorporated, BART-MNLI-e$_p$ and BART-MNLI-e are exactly the same. Our results with BERT-based and XLNet-based models (provided in Appendix G) show similar observations. Next, the BART-MNLI model without fine-tuning on EZ-STANCE performs much worse than the fine-tuned BART-MNLI encoders, particularly for the noun-phrase targets. This result demonstrates the necessity of developing a large dataset for ZSSD, so that the NLI pre-trained knowledge can be fine-tuned and better utilized. Moreover, LLaMA 2 and ChatGPT exhibit much lower performance than fine-tuned transformer-based models, indicating that our dataset is very challenging even for advanced large language models.

6.2 Domain-based Zero-Shot Stance Detection

Domain-based ZSSD (subtask B) focuses on evaluating classifiers using unseen topics from new domains. One domain is selected as the zero-shot domain, and the remaining seven as source domains. Models are trained and validated on source domain data and tested on zero-shot domain data, with eight different zero-shot domain settings in total.

Table 8 shows F1$_{macro}$ scores for various zero-shot domain settings. Models trained on the full mixed-target dataset are evaluated across three settings: 1) the full mixed-target test set; 2) the noun-phrase target-only test set; and 3) the claim target-only test set, denoted as M, N, and C, respectively. First, we notice that models show lower performance when compared with the in-domain subtask A (see results in Table 7). This is because the domain shifts between the training and testing stages introduce additional complexity to the task, making domain-based ZSSD a more challenging ZSSD task. Second, models generally perform worse on the "Covid Epidemic" (CE) and the "Politics" (P) domain, suggesting that these two domains share less domain knowledge with other domains, making them more difficult zero-shot domains. Moreover, we observe that most models show higher performance when predicting stances for the "Rights" (R) and the "Environmental Protection" (EP) domain. BERT-based and XLNet-based models exhibit worse performance than RoBERTa-based models, as illustrated in Appendix G. Class-specific performance for both Subtasks A and B is shown in Appendix H.

Model		CE	WE	EdC	EnC	S	R	EP	P
BiCE	M	.437	.440	.478	.464	.475	.479	.470	.442
	N	.442	.461	.476	.463	.442	.492	.471	.468
	C	.423	.423	.458	.443	.437	.436	.445	.414
CrossNet	M	.501	.513	.514	.500	.509	.546	.521	.512
	N	.475	.492	.480	.483	.473	.508	.512	.486
	C	.518	.525	.530	.513	.505	.534	.509	.516
TGA-Net	M	.568	.569	.583	.580	.572	.635	.566	.577
	N	.531	.554	.556	.551	.563	.588	.565	.555
	C	.589	.568	.600	.602	.572	.631	.565	.582
LLaMA 2	M	.345	.328	.348	.342	.273	.354	.330	.328
	N	.375	.341	.404	.391	.353	.393	.383	.345
	C	.322	.320	.310	.314	.240	.321	.288	.314
ChatGPT	M	.485	.493	.490	.497	.506	.486	.513	.491
	N	.576	.563	.568	.586	.586	.564	.572	.564
	C	.422	.445	.435	.437	.448	.429	.464	.438
RoBERTa	M	.738	.758	.762	.753	.765	.777	.769	.755
	N	.597	.624	.609	.606	.609	.625	.648	.620
	C	.826	.849	.862	.851	.848	.846	.845	.838
RoBERTa-MNLI	M	.760	.773	.777	.776	.778	**.787**	.779	.766
	N	.616	.624	.614	.615	.617	**.642**	.654	.618
	C	.856	.872	.881	.876	.866	.861	.861	.862
RoBERTa-MNLI$_p$	M	.753	.772	.776	.776	.776	**.787**	.778	.767
	N	.597	.617	.603	.621	.608	.638	.642	.623
	C	.855	.876	.884	.877	.868	.866	.863	.863
BART-MNLI	M	.597	.594	.637	.632	.671	.623	.652	.591
	N	.307	.264	.337	.332	.369	.327	.364	.315
	C	.778	.813	.806	.800	.814	.797	.802	.776
BART-MNLI-e	M	.767	.782	**.788**	.776	**.798**	.778	.787	.775
	N	.624	.637	**.625**	.623	**.630**	.612	.665	**.627**
	C	.861	.879	**.890**	**.880**	**.886**	.868	.867	.876
BART-MNLI-e$_p$	M	**.768***	**.789***	.784	**.777***	.792	.783	**.791***	**.777***
	N	**.627***	**.638***	.615	**.625***	.625	.615	**.672***	.624
	C	**.863***	**.890***	.888	.878	.882	**.873***	**.872***	**.877***

Table 8: Subtask B: Comparison of $F1_{macro}$ of models trained and evaluated using 8 zero-shot domain settings (denoted by each column). Models are trained on training set with mixed targets. Test results are denoted as M for mixed, N for noun-phrase, and C for claim targets. $*$: our approach improves the best ZSSD baseline at $p < 0.05$ with paired t-test. Blue, red, and cyan mark the best performance for M, N, and C, respectively.

6.3 EZ-STANCE vs. VAST

We compare EZ-STANCE and VAST to understand which dataset presents more challenges. We consider the following experiments: 1) cross-dataset setting: training our best-performing BART-MNLI-e$_p$ using one dataset and testing the model using the other dataset, and 2) in-dataset setting: training and testing the model on the same dataset. Since VAST is designed for target-based ZSSD and includes only noun-phrase targets, we ensured a fair comparison by utilizing the dataset with noun-phrase targets from subtask A of EZ-STANCE.

Per-class and overall results are shown in Table 9. First, we observe that models show higher performance for the in-dataset setting than the cross-dataset setting. Second, in the in-dataset setting, the model trained on EZ-STANCE (E→E) exhibits much lower performance for the neutral class than its VAST-trained counterpart (V→V). The result demonstrates that data from the neutral class in EZ-STANCE with close semantic correlations between documents and targets are much more challenging than in VAST, where documents and targets are randomly permuted (and do not reflect the natural/real-world data for the neutral class). Third, in the cross-dataset setting, the model trained on VAST performs extremely poorly on the neutral class of the EZ-STANCE test set (V→E), while the model trained on EZ-STANCE show much higher performance on VAST, particularly for the neutral class (E→V). This indicates that EZ-STANCE test set captures more challenging real-world ZSSD data, especially for the neutral category. This reinforces our motivation to create a new, large ZSSD dataset.

Setting	Train/Val	Test	Con	Pro	Neu	All
Cross-dataset	E	V	.600	.593	.309	.501
	V	E	.584	.570	.167	.440
In-data set	E	E	.740	.724	.597	.687
	V	V	.706	.690	.921	.772

Table 9: Cross-dataset and in-dataset performance of BART-MNLI-e$_p$ trained using EZ-STANCE and VAST (denoted as E and V, respectively).

Prompts	Con	Pro	Neu	All
no prompt	.724	.712	.588	.674
The above text entails [target]!	.725	.717	.599	.681
The premise has entailment relation with [target]!	.725	.710	**.602**	.679
This implies the entailment relation with [target]!	.736	.723	.576	.679
The premise has the entailment relation with the hypothesis [target]!	.728	.723	.594	.682
The premise entails the hypothesis [target]!	.730	.702	**.602**	.678
Ours	**.740**	**.724**	.597	**.687**

Table 10: Comparison of $F1_{macro}$ of BART-MNLI-e$_p$ trained using different prompts.

6.4 Impact of Prompt Templates

To assess the efficacy of various prompt templates in our proposed approach, we compare the following prompt settings for noun-phrase targets: 1) applying no prompt to noun-phrase targets; 2) using one of our proposed prompts consistently across all noun-phrase targets; and 3) our approach that randomly assigns each noun-phrase target with one of five distinct prompts. The results with our best-performing BART-MNLI-e$_p$ model are shown in Table 10. We observe that the models trained using our random-prompt approach exhibits better performance than those trained with the singular-prompt approach or no prompts at all.

7 Conclusion

In this paper, we present EZ-STANCE, a large English ZSSD dataset. Compared with existing English ZSSD datasets, our dataset is larger and more challenging. EZ-STANCE covers both noun-phrase targets and claim targets and also comprises two challenging ZSSD subtasks: target-based ZSSD and domain-based ZSSD. We improve the data quality of the neutral class by extracting targets from texts. We evaluate EZ-STANCE on ZSSD baselines and propose to transform ZSSD into the NLI task which outperforms traditional baselines. We hope EZ-STANCE can facilitate future research for varied stance detection tasks.

Limitations

Our EZ-STANCE data is collected from social media. This might be perceived as a drawback as it might not encompass all facets of formal texts that could be found in essays or news comments. In the future, we aim to expand this dataset to include other types of text (e.g., from social media to research articles). Yet, this restriction is not unique to our dataset, but also affects any other datasets that concentrate on social media content.

Ethical Statement

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

Acknowledgements

We thank the National Science Foundation for support from grants 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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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.

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).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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 6298–6304.

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

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085, Toronto, Canada. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. 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.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

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

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.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. Openstance: Real-world zero-shot stance detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning, CoNLL 2022, Abu Dhabi, United Arab Emirates (Hybrid Event), December 7-8, 2022*, pages 314–324. Association for Computational Linguistics.

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, Daijun Ding, and Liwen Jing. 2023. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

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.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for Chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385, Toronto, Canada. 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.

A More Examples of EZ-STANCE

In this section, we show examples of tweets with noun-phrase targets and claim targets for each domain of our EZ-STANCE dataset in Table 11.

B Query Keywords and Data Filtering Strategy

The full keywords set that we used for data crawling is shown in Table 12. We generate the list as follows. First, we use domain names from the *Explore* page of Twitter as keywords for crawling (e.g., sports, education, etc.). This represents our initial set of keywords. After we collect tweets using this initial set, we tokenize the tweets and remove stop words and words with part-of-speeches other than nouns and adjectives, and form words / phrases (from the contiguous words in the text) using the regular expression adjective*noun+. These words/phrases (possibly single words if no adjectives appear in front of the nouns) represent our candidate set of keywords / keyphrases. We then calculate the frequency of these candidates from the tweet collection and rank them in descending order of their frequency. Last, we selected the top-20 keywords / keyphrases as supplementary query keywords / keyphrases to collect tweets in the next iteration. We repeat this process in multiple iterations until we collect enough data.

After we collect the initial keywords set, we perform keyword filtering in the following steps: 1) We manually detect a subset of tweets crawled using each keyword and we remove keywords that are frequently associated with promotional content (e.g., YouTuber, live shopping, etc.), whose main purpose is for product/people promotion instead of addressing controversial topics; 2) Keywords that people predominantly hold single stances on are filtered out, e.g., pollution, crime, delicious food, etc. This is because models would simply learn the correlation between the keywords and the stance and predict stances based solely on keywords instead of the content of tweets and targets.

To ensure the quality of our dataset, we then perform the following preprocessing steps: 1) We remove tweets with less than 20 or more than 150 words. According to our observations, tweets with less than 20 words are either too easy or cannot include enough information to express stances toward multiple targets. Tweets with more than 150 words usually contain links to external content; 2) We remove duplicates and retweets; 3) We keep only tweets in English; 4) We filter out tweets containing advertising contents (e.g., scan the QR code, reply or DM me, sign up, etc.); and 5) We remove emojis and URLs as they may introduce noise. We randomly select around 86 tweets for each keyword, obtaining 6204 tweets for annotation.

C Annotation Instruction

C.1 Noun-phrase Targets

For noun-phrase targets, the annotation takes two steps. For step1, annotators are given the following instructions: *From each tweet, please identify at least 2 noun-phrase targets. Targets should meet the following criteria: 1) Targets should be the principal subject of the tweet rather than minor details; 2) Targets should represent widely discussed topics where different stances are exhibited; 3) Targets where people often express the same stance should be avoided, e.g., violence abuse.* In step 2, we instruct 3 annotators to assign a stance label to each tweet-target pair, using the following instructions: *Imagine yourself as the author of the tweet, please annotate the stance that you would*

CE	Tweet	Cost of living off the scale, country being flooded with migrants, covid scam and jab injuries out there. How much more before the people decide enough is enough.
	N target/Stance	Covid Scams / Against
	C target/ Stance	Skyrocketing living costs and on the other side migrants will come in a lot of amounts so the country's population will increase someday. / Neutral
WE	Tweet	China's economy isn't just doing well. It is increasingly becoming 1 in several categories. Home prices are growing at slow and healthy rates, inflation is normal and healthy and the yuan is solid. The west should be trying to befriend China. Make a friend, not an adversary.
	N target/Stance	China's economy / Favor
	C target/ Stance	The economy of china is decreasing at an alarming rate due to which it's occupied last position in several categories. / Against
EdC	Tweet	To my Twitter pals who are parents in Ontario, trying to deal with homes chooling and work and all the stresses of the pandemic, my God, I don't know how you've managed to pull this off. But you have, even if you're exhausted. And you all rock.
	N target/Stance	home schooling / Against
	C target/ Stance	Parents in Ontario have managed to cope with homeschooling, work, and the pandemic, even if they are exhausted. / Favor
EnC	Tweet	Interviewer: why do you want this position? Me: so I can pay for all the online shopping I did this while being stressed about this interview.
	N target/Stance	online shopping / Favor
	C target/Stance	I do online shopping when I'm stressed. / Neutral
S	Tweet	Dwyane Wade winning an NBA Championship in his 3rd NBA season as the best player on the team .. does not get spoken on enough.
	N target/Stance	Dwyane Wade / Favor
	C target/Stance	Dwyane Wade's success in his 3rd NBA season made him the best player of all times. / Neutral
R	Tweet	The FEUHS Student Government is one with the LGBTQIA community in celebrating the PrideMonth2021 and pursuing equal rights for everyone, regardless of sexual orientation, gender identity, and expression.
	N target/Stance	Equal Rights / Favor
	C target/ Stance	Regardless of sexual orientation, gender identity, or gender expression, the FEUHS Student Government opposes equitable rights for everyone. / Against
EP	Tweet	The Sines coal plant in Portugal has been shut down nine years ahead of schedule, reducing the country s carbon emissions by 12%. A second and final plant is due to close in November which will make Portugal the fourth European country to eliminate.
	N target/Stance	Carbon emissions / Against
	C target/ Stance	Portugal's Sines coal facility was shut down nine years earlier than expected, cutting the nation's carbon emissions by 12 percent. / Favor
P	Tweet	I wish Democrats would play tough and just release an ad that says "GOP loves guns more than our kids." Just show the 234 mass shootings in 2022 and how GOP has obstructed every attempt at gun reform. There's no lie in that claim. At the very least don't call them "rational."
	N target/Stance	GOP / Against
	C target/Stance	The GOP will bring gun reform to stop the mass shootings. / Neutral

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

YouTube shorts, modern history, work from home, herd immunity, living with covid, Fauci, public education, college football, pop culture, war, LGBTQ, environmental awareness, YouTube, career, vaccine, reels, democracy, pop culture, online shopping, hockey, reform, AI assistance writing, working class, election, parenting, global news, China, NBA, sports, student loan, traditional culture, Asian hate, presidential debate, Russia, bully, climate change, medicare, forcing electrical power, Mideast, doctors and patients, anti LGBTQ, post-covid, cooking, Snapchat, EU, presidential election, tictok, pfizer, business, general election, baseketball, prices, Chinese history, insurance, covid conspiracy, live shopping, SAT, Taliban, MLB, baseball, vaccine injury, tiger parents, environmental protection a, gency cultural output, Reels, government, family, new energy, WFH, clean energy, consumption concept, right wing, quality education, world news, stock market, private education, racism, long covid, NFL, vote, negative population growth, youtube, NASA, co-existence with Covid, WWE, DPR, political correctness, world cup, relationship, epidemic prevention, mideast, artificial intelligence, ethical consumption, Garbage classification, arming teachers, force kid to compete, health insurance, media, Negative population growth, terrorism, NATO, population aging, MLB's rule change, technology, wildfire, gun control, gender equality, migrant, doctors and patient, debate, mRNA vaccine, boxing, booster, leftists, republican, life in reels, abortion, teacher carry gun, Disney, overloaded kids, reward unreliable electricity gasoline price, international student, Ukraine, women's football, BLM, DPRK, privacy, shut down coal plants, homeschooling, physical education, men's football, NCAA, security, mask, sealed management, medical insurance, vegetarian, short video, iPhone, Iran, democrat, FDA, mid-term election, livestream shopping, CDC, women's rights, politic, electric vihicles, new york time, Hollywood, immigrant, Metoo, covid-19, equal rights, nuclear energy, mask mandate

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

		# Examples		# Unique			Avg. Length		
		N	C	N	C	T	N	C	T
Covid Epidemic	Train	12,648	19,467	8,506	19,440	6,489	2	18.9	39.4
	Val	1,958	3,753	1,932	3,749	1,251	2.4	19.2	40.4
	Test	2,639	3,819	1,734	3,814	1,273	1.9	19.2	41.8
World Event	Train	12,736	20,025	8,574	19,998	6,675	2	18.9	39.5
	Val	1,996	3,762	1,968	3,755	1,254	2.4	19.2	40.3
	Test	2,286	3,252	1,655	3,252	1,084	1.9	19.3	41.6
Education and Culture	Train	13,054	20,196	8,736	20,169	6,732	2	18.9	39.6
	Val	1,962	3,765	1,940	3,758	1,255	2.3	19.1	39.7
	Test	2,109	3,078	1,515	3,077	1,026	2	19.5	42.2
Entertainment and Consumption	Train	12,760	19,407	8,388	19,386	6,469	2	19.1	40.6
	Val	1,880	3,579	1,850	3,571	1,193	2.4	19.4	40.7
	Test	2,702	4,053	1,949	4,047	1,351	1.9	17.8	35.8
Sports	Train	14,253	20,631	8,838	20,606	6,877	1.9	19	40.4
	Val	1,977	3,747	1,945	3,740	1,249	2.3	19.2	40.5
	Test	1,807	2,661	1,413	2,655	887	2.1	18.4	35.6
Rights	Train	12,619	19,851	8,464	19,824	6,617	2	18.9	39.7
	Val	1,960	3,783	1,936	3,778	1,261	2.4	19.1	40.1
	Test	2,468	3,405	1,793	3,400	1,135	2	19.2	40.5
Environmental Protection	Train	12,989	20,436	8,688	20,406	6,812	2	18.8	39.6
	Val	2,003	3,831	1,978	3,824	1,277	2.3	19.1	39.9
	Test	2,071	2,772	1,519	2,772	924	2.3	19.8	41.9
Politics	Train	12,066	19,419	8,281	19,393	6,473	2	18.9	39.7
	Val	1,846	3,621	1,828	3,617	1,207	2.4	19.3	40.6
	Test	2,890	3,999	2,074	3,995	1,333	1.9	18.9	40.2

Table 13: Data statistics of all 8 dataset splits for subtask B. N, C, and T represent noun-phrase targets, claim targets, and tweets, respectively.

take on this given target as "Favor", "Against", or "Neutral".

C.2 Claim Targets

For claim targets, annotators are provided with the following instructions: *Based on the message that you learned from the tweet, 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 from the claim (against); 3) Based solely on the information from the tweet, we cannot know whether the author definitely supports or opposes the point or message of the claim (neutral).* To make this task more challenging, we establish some extra requirements: First, claims labeled with *favor* must not replicate the tweet verbatim. Second, claims labeled with *against* should not merely negate the tweet content (e.g., adding "not" before verbs). Models could easily detect such linguistic patterns and predict stances without learning the content of tweet-claim pairs.

D Split Method

Initially, we randomly select x% of unique tweets for the training set and the rest as the combination of validation and test set. We then move data with overlapping targets and documents from the mixture of validation and test sets to the training set. After this step, we may introduce some additional overlapping targets during the transaction. This is because the tweets that are moved to the training set may have other noun-phrase targets that overlap with the remaining validation and test set. Therefore we repeat this transferring procedure y times until we do not have any overlapping targets and documents between the training set and the mixture of validation and test set. In our experiments, we use x=40% and y=4, because with these parameters, 66% tweets are split into our final training (similar to VAST). We then perform similar procedures to split validation and the test set. Therefore, the training, validation, and test set do not include overlapping tweets and targets with each other.

E Full Statistics of Subtask B

The statistics of the 8 dataset splits (data from seven domains for training and validation, and the data from the left-out domain as the zero-shot test set) are shown in Table 13.

F Training Details

Our experiments are carried out using an NVIDIA RTX A5000 GPU based on the PyTorch (Paszke et al., 2019). Hyperparameters were fine-tuned using our validation set. The BiCE and CrossNet models were trained using AdamW (Loshchilov

Train/Val	Mixed targets			Noun-phrase targets			Claim targets		
Test	M	N	C	M	N	C	M	N	C
BERT	.760	.636	.823	.408	.633	.288	.653	.320	.816
BERT-MNLI	.772	.625	.847	.420	.638	.305	.679	.338	.851
BERT-MNLI$_p$	.775	.635	.848	.409	.643	.295	-	-	-
XLNet	.775	.642	.842	.427	.651	.310	.670	.305	.842
XLNet-MNLI	.794	.635	.875	.498	.655	.415	.701	.358	.875
XLNet-MNLI$_p$	.799	.641	.878	.452	.663	.345	-	-	-

Table 14: Comparison of F1$_{macro}$ of BERT-based models and XLNet-based models on EZ-STANCE subtask A. M, N, C represent mixed, noun-phrase, and claim targets, respectively.

Model		CE	WE	EdC	EnC	S	R	EP	P
BERT	M	.707	.731	.726	.726	.735	.752	.739	.725
	N	.570	.610	.576	.605	.598	.626	.626	.593
	C	.798	.811	.818	.810	.810	.811	.812	.804
BERT-MNLI	M	.726	.740	.744	.741	.754	.761	.757	.734
	N	.578	.595	.585	.598	.593	.606	.623	.581
	C	.825	.835	.843	.836	.841	.841	.839	.827
BERT-MNLI$_p$	M	.723	.741	.745	.735	.753	.755	.756	.734
	N	.575	.596	.589	.586	.599	.604	.631	.586
	C	.822	.836	.843	.834	.838	.841	.835	.826
XLNet	M	.722	.741	.748	.732	.745	.762	.752	.738
	N	.594	.625	.595	.606	.599	.622	.639	.612
	C	.806	.819	.844	.818	.822	.827	.823	.817
XLNet-MNLI	M	.738	.772	.766	.767	.775	.776	.768	.754
	N	.580	.622	.586	.617	.606	.627	.639	.608
	C	.846	.873	.877	.867	.864	.851	.850	.848
XLNet-MNLI$_p$	M	.749	.765	.766	.760	.775	.767	.768	.762
	N	.600	.609	.594	.605	.605	.604	.629	.620
	C	.850	.868	.873	.863	.865	.854	.855	.852

Table 15: Comparison of $F1_{macro}$ of BERT-based models and XLNet-based models on subtask B. Models are trained and evaluated using datasets for 8 zero-shot domain settings (denoted by each column). Models are trained on the full training set with mixed targets. Test results are denoted as M for mixed targets, N for noun-phrase targets, and C for claim targets.

and Hutter, 2019) as the optimizer with a learning rate of 0.001. Each model was trained for 20 epochs, with each mini-batch of size 128. As for TGA-Net, we adhered to the hyperparameters as recommended in prior research (Allaway and McKeown, 2020). The AdamW optimizer with a learning rate of 2e-5 was utilized for vanilla transformer-based models (BERT, RoBERTa, XLNet), and the NLI pre-trained models (BERT-MNLI$_p$[2], RoBERTa-MNLI$_p$[3], XLNet-MNLI$_p$[4], BART-MNLI-e$_p$[5]), which were fine-tuned for 4 epochs using batch size of 64. The entire training process for each model was completed within 3 hours. Each result is the average of 4 runs with different initializations. For ChatGPT, we utilized the gpt-3.5-turbo-0301 version. We use the following prompt to extract stance predictions using ChatGPT: "*Q: What is the stance of the text '[tweet]' towards the target '[target]'? The answer should be selected from 'Favor', 'Against', or 'None'. A:*" For LLaMA 2, we utilized the Llama-2-13b-chat-hf version with the following prompt: "*Classify the stance that the author of the text takes towards the target into favor, against, or neutral. The answer should only be one of the following three words: 'favor', 'against', or 'neutral'. Don't give further explanation other than one of these three words. Text: "[tweet]". Target: "[target]".*"

G Evaluations on BERT-based and XLNet-based Models for Subtask A and Subtask B

We also evaluate BERT-based models (i.e., BERT, BERT-MNLI, BERT-MNLI$_p$) and XLNet-based models (i.e., XLNet, XLNet-MNLI and XLNet-MNLI$_p$). The results for subtask A and subtask B are shown in Table 14 and Table 15, respectively. First, we can observe that for both subtask A and subtask B, the models with MNLI pre-training (BERT-MNLI, XLNet-MNLI) outperform the corresponding standard BERT and XLNet models. Additionally, when our proposed prompts are applied to noun-phrase targets, BERT-MNLI$_p$ and XLNet-MNLI$_p$ demonstrate enhanced performance compared to BERT-MNLI and XLNet-MNLI, especially in experiments involving noun-phrase targets. This finding underscores the effectiveness of our prompts in improving the MNLI model with BERT and XLNet architectures.

H Class-specific Performance for Subtask A and Subtask B

To establish more comprehensive results, for subtask A and subtask B, we report the class-specific performance. The performance is reported using F1 score for the against (Con), favor (Pro), neutral

[2] https://huggingface.co/textattack/bert-base-uncased-MNLI

[3] https://huggingface.co/textattack/roberta-base-MNLI

[4] https://huggingface.co/textattack/xlnet-base-cased-MNLI

[5] https://huggingface.co/facebook/bart-large-mnli

Train/Val		Mixed targets			Noun-phrase targets			Claim targets		
Test		M	N	C	M	N	C	M	N	C
BiCE	Con	.519	.512	.521	.387	.546	.305	.369	.352	.364
	Pro	.376	.524	.262	.417	.541	.354	.251	.257	.241
	Neu	.509	.393	.568	.390	.500	.344	.237	.238	.343
Cross-Net	Con	..494	.542	.472	.396	.564	.304	.362	.146	.446
	Pro	.509	.549	.485	.426	.571	.365	.497	.495	.512
	Neu	.551	.436	.608	.400	.518	.339	.482	.083	.612
TGA Net	Con	.552	.611	.524	.461	.658	.357	.460	.347	.513
	Pro	.603	.636	.583	.448	.628	.354	.535	.419	.602
	Neu	.614	.491	.674	.364	.534	.280	.531	.189	.672
LLaMA 2	Con	.272	.379	.226	.272	.379	.226	.272	.379	.226
	Pro	.516	.571	.483	.516	.571	.483	.516	**.571**	.483
	Neu	.200	.190	.205	.200	.190	.205	.200	.190	.205
ChatGPT	Con	.556	.653	.511	.556	.653	.511	.556	.653	.511
	Pro	.517	.637	.445	.517	.637	.445	.517	.637	.445
	Neu	.424	.523	.365	.424	.523	.365	.424	.523	.365
BERT	Con	.748	.692	.776	.428	.678	.296	.624	.409	.765
	Pro	.768	.680	.819	.494	.689	.401	.650	.257	.811
	Neu	.765	.535	.875	.302	.533	.165	.686	.295	.872
BERT-MNLI	Con	.774	.678	.818	.459	.682	.351	.667	.386	.823
	Pro	.777	.669	.841	.483	.688	.378	.680	.374	.847
	Neu	.766	.526	.881	.317	.546	.187	.690	.254	.883
BERT-MNLI$_p$	Con	.768	.660	.817	.421	.678	.302	-	-	-
	Pro	.783	.670	.845	.465	.688	.364	-	-	-
	Neu	.775	.575	.882	.341	.563	.220	-	-	-
XLNet	Con	.766	.678	.806	.450	.690	.322	.650	.440	.803
	Pro	.781	.694	.834	.469	.698	.352	.666	.196	.835
	Neu	.776	.553	.886	.362	.565	.256	.694	.280	.887
XLNet-MNLI	Con	.804	.694	.857	.601	.704	.552	.702	**.441**	.859
	Pro	.805	.682	.877	.520	.695	.425	.706	.333	.874
	Neu	.774	.528	.890	.373	.567	.268	.695	.299	.891
XLNet-MNLI$_p$	Con	.807	.702	.859	.533	.703	.454	-	-	-
	Pro	.815	.701	.882	.491	.706	.390	-	-	-
	Neu	.776	.520	.893	.331	.581	.190	-	-	-
RoBERTa	Con	.785	.692	.827	.482	.698	.376	.680	.428	.821
	Pro	.794	.679	.859	.473	.699	.357	.683	.226	.860
	Neu	.773	.547	.888	.345	.571	.219	.699	**.381**	.888
RoBERTa-MNLI	Con	.805	.697	.855	.581	.714	.511	.707	.366	.860
	Pro	.810	.695	.877	.589	.698	.535	.712	.459	.884
	Neu	.775	.535	.894	.404	.564	.308	.699	.193	.897
RoBERTa-MNLI$_p$	Con	.809	.718	.859	.504	.722	.426	-	-	-
	Pro	.810	.700	.881	.617	.689	.590	-	-	-
	Neu	.779	.566	.894	.474	.593	.459	-	-	-
BART-MNLI	Con	.669	.188	.811	**.669**	.188	**.811**	.669	.188	.811
	Pro	.697	.523	.843	**.697**	.523	**.843**	.697	.523	.843
	Neu	.626	.173	.797	**.626**	.173	**.797**	.626	.173	.797
BART-MNLI-e	Con	.823	.708	**.873***	.458	.724	.319	**.726***	.400	**.875***
	Pro	.819	.696	**.890**	.504	.712	.398	**.717**	.454	**.887**
	Neu	.787	**.578***	.886	.391	.588	.286	**.705**	.179	**.903**
BART-MNLI-e$_p$	Con	**.824***	**.724***	.870	.478	**.740***	.362	-	-	-
	Pro	**.825***	**.717***	.888	.494	**.724***	.388	-	-	-
	Neu	**.788**	.567	**.898***	.364	**.597**	.217	-	-	-

Table 16: Comparison of class-specific F1 scores of models on EZ-STANCE subtask A. M, N, C represent mixed, noun-phrase, and claim targets, respectively. The performance is reported using F1 score for the against (Con), favor (Pro), neutral (Neu). $*$: our approach improves the best ZSSD baseline at $p < 0.05$ with paired t-test. Blue, red, and cyan represent best performance in against, favor, and neutral class, respectively.

Model		CE	WE	EdC	EnC	S	R	EP	P
BiCE	M	.514	.544	.521	.498	.443	.575	.457	.558
	N	.492	.553	.502	.436	.212	.620	.432	.607
	C	.524	.538	.528	.520	.477	.545	.467	.523
CrossNet	M	.500	.533	.477	.478	.452	.555	.504	.549
	N	.541	.588	.487	.472	.309	.656	.535	.627
	C	.470	.491	.471	.480	.479	.454	.485	.474
TGA-Net	M	.553	.550	.536	.545	.497	.636	.519	.584
	N	.598	.635	.585	.545	.455	.709	.551	.662
	C	.521	.472	.503	.545	.507	.577	.502	.510
LLaMA 2	M	.329	.315	.288	.248	.119	.342	.226	.326
	N	.407	.358	.371	.359	.276	.399	.286	.383
	C	.277	.286	.240	.192	.082	.303	.194	.282
ChatGPT	M	.567	.578	.559	.523	.474	.592	.586	.603
	N	.640	.632	.614	.594	.520	.651	.577	.685
	C	.516	.536	.525	.482	.460	.547	.591	.537
BERT	M	.695	.718	.711	.706	.722	.731	.739	.732
	N	.633	.673	.630	.608	.506	.721	.673	.708
	C	.741	.756	.761	.755	.770	.738	.778	.755
BERT-MNLI	M	.727	.738	.746	.736	.765	.755	.768	.748
	N	.654	.662	.641	.614	.513	.710	.682	.701
	C	.786	.802	.811	.800	.823	.795	.820	.794
BERT-MNLI$_p$	M	.721	.747	.743	.733	.765	.750	.763	.745
	N	.641	.678	.648	.608	.517	.699	.672	.696
	C	.781	.806	.807	.798	.821	.794	.817	.793
XLNet	M	.720	.740	.752	.713	.740	.757	.759	.742
	N	.673	.697	.651	.614	.493	.736	.694	.711
	C	.758	.777	.815	.768	.793	.775	.795	.772
XLNet-MNLI	M	.741	.781	.774	.769	.794	.779	.783	.766
	N	.640	.690	.640	.625	.511	.739	.691	.711
	C	.818	.857	.856	.844	.858	.815	.835	.818
XLNet-MNLI$_p$	M	.754	.779	.779	.766	.796	.774	.785	.775
	N	.664	.694	.648	.625	.522	.727	.684	.730
	C	.824	.852	.853	.838	.856	.816	.841	.818
RoBERTa	M	.745	.766	.769	.752	.768	.775	.780	.765
	N	.694	.703	.656	.625	.522	.744	.702	.731
	C	.789	.818	.839	.819	.826	.802	.824	.798
RoBERTa-MNLI	M	.769	.779	.791	.777	.793	.788	.797	.785
	N	.688	.700	.673	.626	.540	.748	.702	.726
	C	.832	.847	.864	.853	.857	.823	.848	.840
RoBERTa-MNLI$_p$	M	.764	.787	.790	.779	.794	.794	.795	.789
	N	.687	.708	.662	.623	.534	.748	.691	.730
	C	.827	.855	.867	.854	.860	.834	.850	.842
BART-MNLI	M	.574	.587	.629	.627	.735	.568	.654	.562
	N	.224	.193	.200	.200	.290	.225	.232	.236
	C	.761	.794	.797	.776	.822	.760	.803	.754
BART-MNLI-e	M	.779	.798	.804	.780	.820	.786	.809	.799
	N	.692	.720	.691	.638	.542	.723	.730	.738
	C	.844	.862	.872	.859	.876	.837	.855	.857
BART-MNLI-e$_p$	M	.786*	.807*	.801	.775	.812	.798*	.812*	.801*
	N	.703*	.723*	.682	.626	.543*	.738	.722	.738*
	C	.849*	.876*	.871	.855	.872	.849*	.862*	.857*

Table 17: Comparison of F1 for the "**against**" class of different models trained on mixed targets for 8 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. $*$: our approach improves the best ZSSD baseline at $p < 0.05$ with paired t-test. Blue, red, and cyan represent best performance in mixed targets, noun-phrase targets, and claim targets, respectively.

(Neu). Results are shown in Table 16 (subtask A), Table 17, Table 18, and Table 19 (subtask B).

Model		CE	WE	EdC	EnC	S	R	EP	P
BiCE	M	.300	.257	.424	.390	.430	.371	.478	.254
	N	.420	.397	.580	.478	.538	.602	.642	.445
	C	.190	.153	.284	.245	.295	.180	.324	.139
CrossNet	M	.459	.446	.542	.464	.499	.545	.557	.455
	N	.420	.397	.580	.478	.538	.602	.642	.445
	C	.478	.470	.508	.449	.461	.499	.461	.459
TGA-Net	M	.571	.539	.610	.575	.610	.643	.605	.560
	N	.564	.462	.607	.560	.647	.702	.655	.542
	C	.576	.570	.611	.583	.580	.585	.556	.570
LLaMA 2	M	.489	.459	.535	.520	.561	.547	.582	.486
	N	.518	.433	.609	.569	.642	.629	.681	.489
	C	.466	.478	.475	.483	.496	.476	.490	.483
ChatGPT	M	.483	.444	.528	.491	.554	.556	.615	.510
	N	.590	.518	.633	.607	.641	.672	.713	.576
	C	.408	.399	.443	.406	.475	.459	.517	.467
BERT	M	.712	.730	.748	.724	.737	.768	.788	.715
	N	.586	.571	.666	.611	.644	.728	.752	.602
	C	.792	.808	.820	.800	.812	.803	.825	.791
BERT-MNLI	M	.735	.738	.755	.750	.746	.768	.806	.732
	N	.593	.545	.654	.627	.623	.699	.760	.587
	C	.825	.830	.844	.838	.843	.829	.856	.823
BERT-MNLI$_p$	M	.732	.736	.761	.748	.753	.763	.805	.735
	N	.592	.547	.661	.629	.651	.686	.763	.592
	C	.822	.829	.845	.836	.839	.828	.848	.825
XLNet	M	.729	.732	.759	.740	.748	.771	.797	.734
	N	.614	.582	.670	.623	.668	.730	.765	.608
	C	.801	.806	.838	.813	.817	.807	.833	.812
XLNet-MNLI	M	.751	.771	.788	.776	.776	.787	.813	.758
	N	.598	.581	.682	.641	.670	.727	.760	.614
	C	.848	.867	.885	.869	.867	.840	.869	.855
XLNet-MNLI$_p$	M	.765	.770	.784	.769	.778	.774	.818	.772
	N	.626	.583	.684	.630	.678	.696	.768	.629
	C	.850	.865	.877	.865	.867	.838	.873	.860
RoBERTa	M	.748	.760	.779	.757	.763	.791	.818	.755
	N	.637	.594	.676	.617	.645	.742	.774	.629
	C	.821	.848	.865	.851	.853	.835	.865	.839
RoBERTa-MNLI	M	.769	.774	.789	.788	.776	.804	.821	.770
	N	.622	.583	.679	.652	.658	.744	.769	.622
	C	.858	.870	.885	.882	.871	.856	.876	.864
RoBERTa-MNLI$_p$	M	.766	.773	.792	.786	.778	.800	.829	.773
	N	.625	.584	.686	.654	.665	.734	.784	.631
	C	.858	.874	.889	.880	.873	.856	.879	.867
BART-MNLI	M	.641	.623	.684	.670	.711	.689	.723	.632
	N	.462	.372	.530	.502	.602	.562	.636	.436
	C	.808	.854	.843	.836	.837	.830	.837	.822
BART-MNLI-e	M	.782	.786	.808	.799	.790	.789	.830	.787
	N	.650	.618	.707	.677	.666	.716	.778	.635
	C	.866	.880	.898	.887	.890	.857	.886	.883
BART-MNLI-e$_p$	M	.783*	.796*	.802	.797	.790*	.795	.832*	.785
	N	.658*	.613	.703	.669	.672	.717	.777	.638
	C	.864	.894*	.894	.886	.886	.865*	.890*	.882

Table 18: Comparison of F1 for the "**favor**" class of different models trained on mixed targets for 8 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. $*$: our approach improves the best ZSSD baseline at $p < 0.05$ with paired t-test. Blue, red, and cyan represent best performance in against, favor, and neutral class, respectively.

Model		CE	WE	EdC	EnC	S	R	EP	P
BiCE	M	.496	.520	.489	.505	.552	.491	.474	.513
	N	.465	.491	.371	.498	.573	.265	.360	.387
	C	.556	.577	.561	.563	.539	.583	.543	.581
CrossNet	M	.544	.559	.524	.558	.575	.539	.501	.531
	N	.465	.491	.371	.498	.573	.265	.360	.387
	C	.606	.614	.610	.609	.575	.650	.582	.617
TGA-Net	M	.579	.618	.602	.621	.610	.625	.573	.588
	N	.431	.563	.475	.549	.586	.354	.488	.461
	C	.670	.663	.687	.679	.631	.729	.637	.664
LLaMA 2	M	.215	.210	.221	.258	.141	.174	.182	.172
	N	.199	.231	.234	.245	.140	.153	.183	.163
	C	.224	.195	.214	.267	.142	.184	.182	.178
ChatGPT	M	.405	.457	.383	.478	.491	.311	.339	.360
	N	.498	.538	.455	.558	.598	.367	.425	.432
	C	.342	.399	.336	.423	.411	.281	.286	.312
BERT	M	.713	.745	.720	.750	.746	.758	.689	.727
	N	.492	.586	.431	.595	.643	.429	.454	.469
	C	.860	.869	.872	.874	.847	.891	.832	.867
BERT-MNLI	M	.715	.742	.729	.736	.751	.760	.698	.724
	N	.487	.578	.459	.552	.643	.408	.427	.455
	C	.863	.873	.875	.869	.858	.899	.842	.864
BERT-MNLI$_p$	M	.715	.740	.731	.724	.743	.751	.699	.721
	N	.492	.562	.458	.521	.628	.428	.459	.470
	C	.864	.873	.876	.867	.854	.901	.842	.860
XLNet	M	.716	.751	.733	.743	.748	.760	.699	.737
	N	.495	.596	.465	.580	.634	.399	.459	.518
	C	.857	.872	.879	.874	.854	.898	.842	.868
XLNet-MNLI	M	.721	.763	.737	.755	.756	.761	.708	.737
	N	.503	.596	.436	.585	.636	.416	.465	.498
	C	.871	.895	.890	.889	.867	.899	.847	.873
XLNet-MNLI$_p$	M	.727	.747	.734	.745	.750	.754	.702	.740
	N	.510	.552	.451	.559	.614	.389	.435	.501
	C	.875	.889	.888	.886	.872	.908	.850	.878
RoBERTa	M	.719	.749	.739	.750	.763	.764	.707	.744
	N	.461	.575	.496	.577	.660	.388	.469	.500
	C	.868	.881	.882	.884	.866	.902	.845	.877
RoBERTa-MNLI	M	.741	.766	.750	.764	.764	.770	.719	.742
	N	.540	.588	.491	.598	.652	.434	.492	.506
	C	.879	.899	.893	.894	.870	.905	.859	.881
RoBERTa-MNLI$_p$	M	.729	.757	.744	.764	.755	.768	.710	.740
	N	.479	.560	.459	.597	.625	.431	.451	.507
	C	.880	.898	.896	.896	.871	.910	.861	.879
BART-MNLI	M	.576	.572	.599	.600	.567	.613	.580	.579
	N	.235	.226	.282	.295	.214	.195	.224	.272
	C	.764	.792	.778	.789	.783	.801	.766	.751
BART-MNLI-e	M	.740	.760	.752	.752	.785	.759	.721	.748
	N	.530	.574	.477	.553	.681	.398	.489	.507
	C	.879	.894	.899	.893	.891	.908	.859	.887
BART-MNLI-e$_p$	M	.736	.764	.747	.758	.775	.757	.729*	.746
	N	.521	.577	.461	.580	.659	.389	.518*	.497
	C	.876	.901	.898	.892	.887	.905	.865	.891*

Table 19: Comparison of F1 for the "**neutral**" class of different models trained on mixed targets for 8 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. $*$: our approach improves the best ZSSD baseline at $p < 0.05$ with paired t-test.