Guiding Computational Stance Detection with Expanded Stance Triangle Framework

Zhengyuan Liu[†], **Yong Keong Yap**[‡], **Hai Leong Chieu**[‡], **Nancy F. Chen**[†]
[†]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[‡]DSO National Laboratories, Singapore
{liu_zhengyuan,nfychen}@i2r.a-star.edu.sg
{yyongkeo,chaileon}@dso.org.sg

Abstract

Stance detection determines whether the author of a piece of text is in favor of, against, or neutral towards a specified target, and can be used to gain valuable insights into social media. The ubiquitous indirect referral of targets makes this task challenging, as it requires computational solutions to model semantic features and infer the corresponding implications from a literal statement. Moreover, the limited amount of available training data leads to subpar performance in out-of-domain and cross-target scenarios, as data-driven approaches are prone to rely on superficial and domain-specific features. In this work, we decompose the stance detection task from a linguistic perspective, and investigate key components and inference paths in this task. The stance triangle is a generic linguistic framework previously proposed to describe the fundamental ways people express their stance. We further expand it by characterizing the relationship between explicit and implicit objects. We then use the framework to extend one single training corpus with additional annotation. Experimental results show that strategically-enriched data can significantly improve the performance on out-of-domain and cross-target evaluation.

1 Introduction

Stance (and its variant stancetaking) is a concept defined as a linguistically articulated form of social action whose meaning is construed within language, interaction, and sociocultural value (Biber and Finegan, 1988; Agha, 2003; Du Bois, 2007; Kiesling, 2022). Its subject can be the speaker in a conversation or the author of a social media post, and its object can be in the form of an entity, concept, idea, event, or claim.[1]

The stance detection task in natural language processing aims to predict the stance of a piece of text toward specified targets. Stance detection is commonly formulated as a classification problem (Küçük and Can, 2020), and is often applied to analyzing online user-generated content such as Twitter and Facebook posts (Mohammad et al., 2016; Li et al., 2021). When given the text and one specified target (i.e., stance object), a classifier is used to predict a categorical label (e.g., *Favor*, *Against*, *None*). Along with social networking platforms' growing impact on our lives, stance detection is crucial for various downstream tasks such as fact verification and rumor detection, with wide applications including analyzing user feedback and political opinions (Glandt et al., 2021). For example, during the pandemic of COVID-19, it was essential to understand the public's opinion on various initiatives and concerns, such as getting booster vaccinations and wearing facial masks. The insight from stance analysis could help public health organizations better estimate the expected efficacy of their mandates, as well as proactively detect pandemic fatigue before it leads to a serious resurgence of the virus.

While state-of-the-art results have been achieved on text classification by adopting data-driven neural approaches, especially utilizing recent large-scale language backbones (Devlin et al., 2019; Liu et al., 2019), stance detection remains challenging;

Text: Service was slow, but the people were friendly.	
Aspect: "Service"	**Sentiment**: Negative
Aspect: "people"	**Sentiment**: Positive
Text: I believe in SCIENCE. I wear a mask for YOUR PROTECTION.	
Target: "wear a mask"	**Stance**: Favor ✔
Target: "Dr. Fauci"	**Stance**: Favor ✔
Target: "no mask activity"	**Stance**: Favor ✗
Target: "CD Disk"	**Stance**: Favor ✗

Table 1: Two examples of aspect-level sentiment analysis and target-aware stance detection. The incorrect label prediction is highlighted in red. The target with implicit mention is highlighted in blue.

[1]There are various definitions of stance and stancetaking in pragmatics and sociolinguistics field. In this article, we follow the generic "stance act" defined by Du Bois (2007).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3987–4001

there is a substantial gap between human and machine performance.[2] One challenge comes from the ubiquitous indirect referral of targeted stance objects. When interacting socially online, people express their subjective attitude with brevity and variety: they often do not directly mention the final target, but mention its related entities, events, concepts, or claims. As examples shown in Table 1, unlike aspect-based sentiment analysis, where aspect terms are usually explicitly stated in sentences, targets specified for stance labeling can be flexibly assigned. For instance, in a tweet about COVID-19, while "Dr. Fauci" is not mentioned, one can infer that the user stands for him from the support of "wearing a mask" and "science". Therefore, target-aware context understanding requires capturing the relationship of explicitly-mentioned objects and various targets, but existing models lack such capability.

Another challenge stems from limited annotated data for stance detection. When training on a corpus constructed with a small number of targets from a single domain, data-driven approaches cannot generalize well on out-of-domain samples and unseen targets (Allaway and Mckeown, 2020; Kaushal et al., 2021). Meanwhile, due to low data diversity and the spurious correlation caused by single target labeling, models are prone to over-fit on superficial and biased features (e.g., sentiment-related lexicon). The strong baselines are observed to solely rely on the input text (e.g., tweets) but neglect the specified target (Ghosh et al., 2019; Kaushal et al., 2021), and fail to make correct predictions when we change the targeted object. As shown in Figure 1, the classifier always produces the same output *Favor*, even when irrelevant targets such as "CD Disk" are indicated.

In this work, we investigate solutions for the aforementioned challenges from a linguistic perspective. The pragmatic and linguistics studies provide us with detailed theories of how humans perform stancetaking (Du Bois and Kärkkäinen, 2012; Kiesling et al., 2018), and help us identify the key components and inference paths for stance analysis. The "Stance Triangle" (Du Bois, 2007) is one of the most influential and generic linguistic frameworks.

[2]Recent empirical evaluation studies show that large language models (LLMs) can provide reasonable results on various NLP tasks including stance detection. However, the performance of adopting zero-shot and few-shot inference with LLMs is still lower than task-specific fine-tuned approaches, and LLMs require great amounts of computational resources (Ziems et al., 2023; Bang et al., 2023).

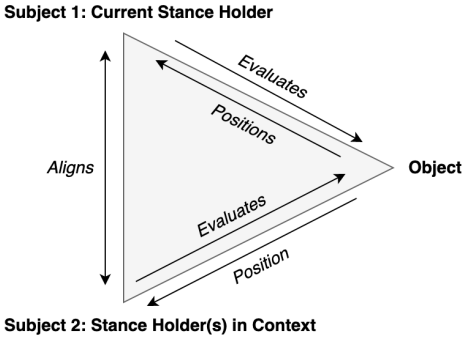

Figure 1: The stance triangle framework proposed by Du Bois (2007). Vertices denote the three basic components. Edges denote expression act types.

As shown in Figure 1, it presents three stancetaking acts: a subject (i.e., the stance holder) evaluates an object, positions themselves and others, and aligns with other subjects. While this model covers the important aspects of stancetaking, its broadness leaves the operationalization of stance in practical use cases under-specified (Kiesling, 2022). Regarding stance analysis of social networking platforms, modeling the implication of targets is important, but it is not well-formulated in the triangle framework. Therefore, we expand it by delineating the relationship between explicit and implicit objects, and outline two paths to complete the human-like inference. Aside from using the expanded framework for qualitative analysis, we further utilize it for strategic annotation enrichment, which shows strong potential to improve the robustness and generality of data-driven approaches. In summary, our contributions of this work are as follows:

- We make the first attempt to expand the linguistic framework "stance triangle" for improving computational stance detection, by characterizing the relationship and labels of explicit and implicit objects.
- We conduct qualitative analysis following the expanded framework on tweet stance detection, and outline the primary aspects and inference paths.
- We leverage the proposed framework to enrich the annotation of a single-domain corpus, and empirically demonstrate its effectiveness in improving the performance of out-of-domain and cross-target generalization.

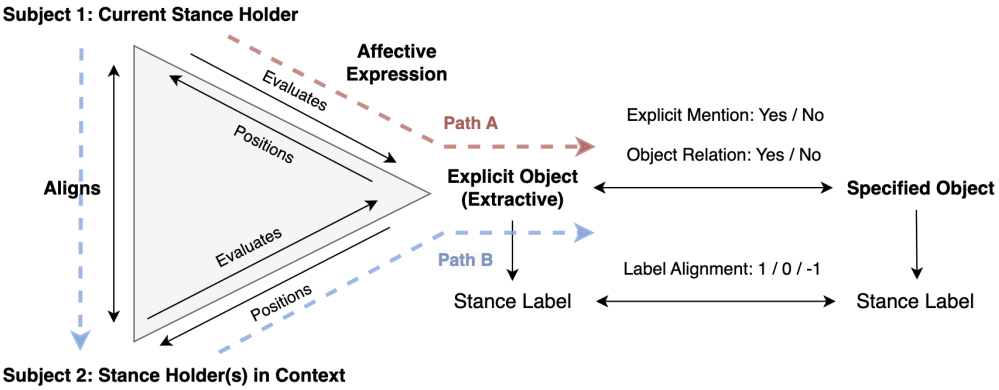

Figure 2: Overview of our proposed framework expanded on the stance triangle model. The two paths of stancetaking flow are shown by dotted arrow-line in pink and blue color.

2 The Stance Triangle Framework

In linguistics, stance is originally defined as "the overt expression of an author's or speaker's attitudes, feelings, judgments, or commitment concerning the message" (Biber and Finegan, 1988). Sociolinguists further emphasized inter-subjective relations, which refers to how speakers position themselves with their own or other people's views, attitudes, and claims toward particular objects or ideas in ongoing communicative interaction (Haddington et al., 2004; Jaffe et al., 2009). In this way, the core concept of stance is embedded with inherently dialogic and inter-subjective characteristics. When other interlocutors in the context indicate a stance on objects, the current speaker may take the same polarity or against their stance in various ways (Du Bois, 2007).

Given that stance is firmly grounded in the communicative interaction and stancetaking is crucial for the social construction of meaning in different discourses, Du Bois (2007) proposed a holistic analytic framework named "**Stance Triangle**", which describes the communication situation in which two speakers create intersubjectivity through their evaluation of an object.

As shown in Figure 1, in the stance triangle, the three key components regarding stancetaking are located at the vertices, namely the current stance holder (**Subject 1**), what the stance is about (**Object**), and the other stance holders in context (**Subject 2**). The edges are then categorized into three types to reveal expression acts among interlocutors and objects. The evaluation and positioning are both **Affective Expression** (Du Bois and Kärkkäinen, 2012; Kiesling et al., 2018). **Evaluation** is tied to affect (Du Bois and Kärkkäinen, 2012). It refers to a person expressing an overt emotional reaction to or displaying an affective orientation toward the object. **Positioning** often occurs as a consequence of evaluation. The positioned subject is the one who makes the evaluation toward a specific object. By invoking an evaluation, the subject presents himself/herself as taking a particular affective orientation toward an object via the act of stancetaking. **Alignment** is used to highlight the similarities and differences of interlocutors' stance. In communicative interaction, one could find his/her own stance is compared and contrasted with others. As communication goes on, interlocutors may line up affective stances on the shared object. Alternatively, they negotiate their stance and stay differential. Alignment thus becomes an essential part of intersubjectivity in dynamic dialogical interaction.

The stance triangle provides insights into the dialogic nature of stance and serves as an analytical framework to understand how stance is taken in interactions. There are many qualitative studies that are inspired by this framework and applied the fundamental theory to analyze various resources, such as story narration (Bohmann and Ahlers, 2022), and social media text (Simaki et al., 2018). While previous work demonstrates the success of the stance triangle in linguistics and social constructionism, to the best of our knowledge, few works use the linguistic framework to facilitate computational stance detection.

3 Our Proposed Linguistic Framework

While this triangle model describes the important aspects of stancetaking, its broadness leaves a few limitations when we adopt it to analyze online user-generated content. First, the triangle model mainly

Example A: I believe in SCIENCE. I wear a mask for YOUR PROTECTION.			
Explicit Object: "wear a mask" / "science"		**Stance Label**: Favor	
Specified Target: "Dr. Fauci"	**Explicit Mention**: No	**Label Alignment**: 1	**Stance Label**: Favor
Specified Target: "covid is a hoax"	**Explicit Mention**: No	**Label Alignment**: -1	**Stance Label**: Against
Specified Target: "CD Disk"	**Explicit Mention**: No	**Label Alignment**: 0	**Stance Label**: None
Example B: So can unborn children have rights now?			
Explicit Object: "unborn children"		**Stance Label**: Favor	
Specified Target: "fetus"	**Explicit Mention**: No	**Label Alignment**: 1	**Stance Label**: Favor
Specified Target: "Abortion"	**Explicit Mention**: No	**Label Alignment**: -1	**Stance Label**: Against
Specified Target: "Trump"	**Explicit Mention**: No	**Label Alignment**: 0	**Stance Label**: None

Table 2: Two Path-A examples decomposed and annotated based on our expanded stance triangle framework. The original specified target and its label are highlighted in blue.

Example A: Greater is He who is in you than he who is in the world. - 1 John 4:4			
Explicit Object: "1 John 4:4" (Quotation)		**Stance Label**: Favor	
Specified Target: "God"	**Explicit Mention**: No	**Label Alignment**: 1	**Stance Label**: Favor
Specified Target: "Atheism"	**Explicit Mention**: No	**Label Alignment**: -1	**Stance Label**: Against
Example B: We remind ourselves that love means to be willing to give until it hurts - Mother Teresa			
Explicit Object: "Mother Teresa" (Quotation)		**Stance Label**: Favor	
Specified Target: "the unborn"	**Explicit Mention**: No	**Label Alignment**: 1	**Stance Label**: Favor
Specified Target: "Abortion"	**Explicit Mention**: No	**Label Alignment**: -1	**Stance Label**: Against

Table 3: Two Path-B examples decomposed and annotated based on our expanded stance triangle framework. The original specified target and its label are highlighted in blue.

focuses on dialogic interactions with one shared and explicit object, and does not consider the multiplicity of objects, coreference, and specific quotations on social networking platforms. Meanwhile, the ubiquitous indirect referral of stance targets, and expressions of sarcasm, metaphor, and humor might often require multi-hop reasoning. Moreover, the stance triangle model only presents the interaction of subject-subject and subject-object, and overlooks the object-object relationship. Therefore, we expand the *object* component to two subcomponents: explicit object and specified object, as shown in Figure 2.

Explicit Object denotes the explicitly-mentioned object in the text, which the speaker poses a stance on. Its stance label can be obtained from the literal affective expression or alignment between stance holders. Upon this definition, the explicit object can be obtained in an extractive manner from the context, and one piece of text may contain multiple explicit objects, such as the "wear a mask" and "science" of Example A shown in Table 2.

Specified Object denotes the target for predicting the stance label. How the specified object links to an explicit object determines the alignment or dis-alignment of their stance labels.

Explicit Mention denotes whether the specified target is an explicit object in the context.

Object Relation indicates whether the stance label of a specified target can be inferred from the explicit object. For instance, regarding the irrelevant target "CD Disk", its object relation to the explicit objects "wear a mask" and "science" is *No*, and that of "Dr. Fauci" is *Yes*.

Label Alignment indicates the relationship between a specified target and the explicit object in stance labeling. A value of 1 means they share the same polarity (e.g., *Favor* vs. *Favor*), and -1 means the opposite polarity (e.g., *Favor* vs. *Against*). Moreover, we add the categorical value 0 when they do not have any stance label correlation (i.e., object relation value is *No*). Therefore, the label alignment can also be used to describe the object relation.

We then outline the two inference paths of the stancetaking flow: (1) As **Path-A** shown in Figure 2, the speaker poses an affective expression on the object from a first-person's perspective. For instance, as Example A shown in Table 2, the speaker stands for wearing a face mask to protect others in the COVID-19 pandemic. When the specified target is "covid is a hoax" which is an implicit object, one can infer that it is related to the explicit object "wearing a mask", and their label alignment is -1 (i.e., opposite). Thus the stance label of "covid is a hoax" is *Against*.

Corpus	Targets for Stance Labeling	Train	Valid	Test
SemEval-16 Task-6 A	Atheism, Climate Change, Feminist Movement, Hillary Clinton, Legalization of Abortion	2,914	-	1,249
SemEval-16 Task-6 B	Donald Trump (for zero-shot evaluation)	-	-	707
P-Stance	Donald Trump, Joe Biden, Bernie Sanders	19,228	2,462	2,374
VAST	Various Targets by Human Annotation	13,477	2,062	3,006
Tweet-COVID	Keeping Schools Closed, Dr. Fauci, Stay at Home Orders, Wearing a Face Mask	4,533	800	800

Table 4: Statistics of the collected stance detection datasets for model training and evaluation.

(2) As **Path-B** shown in Figure 2, the current stance holder may align or dis-align with other stance holders. For Example A shown in Table 3, the speaker quoted a sentence from one chapter "John 4:4" of the Bible. This presents the speaker's belief in God, and one can infer that the label of the specified target "Atheism" is *Against*. Moreover, regarding stance analysis of online social networking platforms such as Twitter, the alignment act can be extended to include quotations, re-tweet behavior, and the 'Like' button.

The expanded linguistic framework describes the key components, expression acts, and inference flows of stance detection, and it is helpful for qualitative and quantitative analysis, especially for the implicitly-mentioned stance objects. More importantly, the expanded framework sheds light on the challenging parts of computational stance detection. For instance, some domain knowledge is necessary for reasoning the label alignment between explicit objects and specified targets, and current tweet-related datasets do not provide particular labeling of re-tweet and quotations. This framework paves the way for further research extension.

4 Theory-inspired Practice: Strategic Annotation Enrichment

Various corpora with target-aware stance annotation are constructed to facilitate computational solutions for stance detection (Mohammad et al., 2016; Allaway and Mckeown, 2020; Li et al., 2021; Glandt et al., 2021). However, most of them only adopt a simple annotation scheme, where a single target and its corresponding label are provided. Some recent datasets adopt a multi-target annotation (Kaushal et al., 2021), but the paired target number is limited.

According to our linguistic framework, modeling the implication of a specified object is important. Therefore, we enrich the annotation of one corpus from a single domain by adding multi-target stance labeling on explicit and implicit objects. We select the data from SemEval2016 Task-6 A "tweet stance detection" (Mohammad et al., 2016) as it serves as a benchmark in many previous works (Küçük and Can, 2020). As shown in Table 4, it is built on tweets about a set of politics-relevant targets (e.g., politicians, feminism movement, climate change), and each sample only has one specified target with a human-annotated stance label.

We first obtain a sample subset where the specified target is not explicitly mentioned in the text. Next, to obtain explicit objects in an extractive manner, we apply an off-the-shelf constituency parsing model,[3] and collect all noun phrases in the constituency tree. To reduce extraction noise, we filter out the noun-phrase candidates with some criteria (e.g., being not near the verbs in the sentence, being shorter than 4 characters, and being started with hashtag token and "@user").

Then linguistic annotators are invited to label the stance polarity on the explicit objects. To reduce superficial target-related patterns and biases from single target labeling, and emphasize object-object relationship, here we propose and adopt an adversarial multi-target strategy, namely selecting the explicit object that shows a stance dis-alignment to the specified target (e.g., "unborn children" and "abortion" of Example B in Table 2). This adversarial multi-target labeling can encourage models to condition their prediction more on the given target, as well as learn some correlation between explicit and implicit objects. We obtain 1,500 paired samples, where the original training size is 2,914. Note that we do not introduce any new data to the training set, but enrich the existing corpus. Similar to previous work (Mohammad et al., 2016), our four linguistic annotators participate in the enrichment task (see Appendix Table 9 for more details), and the Cohen's Kappa score calculated for inter-annotator agreement is 0.79 for stance labeling, and according to Uebersax (1982), this score represents a reasonable agreement level.

[3] https://demo.allennlp.org/constituency-parsing

5 Experiments on Computational Stance Detection

5.1 Task Definition

Given $x = \{w_1, w_2, .., w_n\}$ (n denotes the token number) as one input text, and $t = \{t_1, t_2, .., t_m\}$ (m denotes the token number) as the target, the stance detection model is to predict the classification label (e.g., *Favor*, *Against*, *None*). In our experimental setting, we use the 3-class scheme, as the '*None*' label is necessary for practical use cases. Note that in some stance detection corpora, they introduce '*Neutral*' as the third label. To uniform the 3-class labeling for extensive evaluation, we merge '*None*' and '*Neutral*' as one category.

5.2 Target-Aware Classification

The large-scale pre-trained language models yield state-of-the-art performance in text classification and stance detection (Devlin et al., 2019; Kaushal et al., 2021). Here we use a Transformer neural network (Vaswani et al., 2017) as the base architecture, and leverage prior language knowledge by initializing it with the *RoBERTa* (Liu et al., 2019).

Target-Aware Encoding Since predicting the stance of an input text is dependent on the specified target, previous studies show that conditioning the contextualized representation on the target provides substantial improvements (Augenstein et al., 2016; Du et al., 2017; Allaway and Mckeown, 2020), thus we concatenate the input text x and the specified target t as one sequence, and use the language backbone to encode it. The input for encoder is formulated as "<s> t </s> <s> x </s>". Then the pooled output of the final layer hidden representation of the first "<s>" $v_{\text{enc}} \in \mathbb{R}^E$ (where E is the dimension size) is used as the encoded representation of the target-conditioned input.[4]

Label Prediction To predict the stance label, we feed the encoded representation v_{enc} to a fully-connected layer and a softmax function to compute the output probabilities:

$$y^{pred} = \text{softmax}(W^{'} v_{\text{enc}} + b^{'}) \quad (1)$$

where $W^{'}$ and $b^{'}$ are learnable parameters, and the cross-entropy between gold label y^{gold} and model prediction y^{pred} is minimized as the training loss.

[4]The special tokens vary in different language backbones. For BERT-based models (Devlin et al., 2019), <s> and </s> are replaced with [CLS] and [SEP], respectively.

Test Set	In-Domain	Cross-Target
SemEval16 Task-6 A	Yes	No
SemEval16 Task-6 B	Yes	Yes
P-Stance	Yes	Yes
VAST	No	Yes
Tweet-COVID	No	Yes

Table 5: Evaluation setting on different test sets. ***SemEval16 Task-6 A*** is used as the single-domain training corpus for cross-domain and cross-target evaluation.

5.3 Experimental Corpora

We select several representative stance detection datasets for extensive evaluation, including SemEval-16 Task-6 A and Task-6 B (Mohammad et al., 2016), P-Stance (Li et al., 2021), VAST (Allaway and Mckeown, 2020), and Tweet-COVID (Glandt et al., 2021). We use their official train, validation, and test splits. The detailed statistics of these datasets are shown in Table 4. To uniform the label of *None* and *Neutral* from the data perspective, we extend the *None* subset with 20% size of the training data, by extracting irrelevant objects from random samples, as previous contrastive learning study (Gao et al., 2021).

In our experiments, models are basically trained on a single-domain corpus (*SemEval-16 Task-6 A*), and evaluated on multiple test sets. As shown in Table 4, since there are only 5 targets in the single training set of politics-related tweets, testing on different corpora will build the in-domain, out-of-domain, and cross-target evaluation settings (Küçük and Can, 2020). As shown in Table 5, *SemEval-16 Task-6 B* contains unseen target "Donald Trump", which is used to test the cross-target generalization, and testing on *Tweet-COVID* is both out-of-domain and cross-target.

Moreover, stance label distributions of different targets in our tested benchmark corpora are relatively balanced, and this mitigates the concern of model's over-fitting on "target-related patterns" at the evaluation stage.

5.4 Training Configuration

Models are implemented with Pytorch and Hugging Face Transformers[5]. For fine-tuning on stance detection task, we train the language backbone *RoBERTa-base* (Liu et al., 2019) with the AdamW optimizer (Kingma and Ba, 2015) and batch size 32. Initial learning rates are all set at $2e^{-5}$, and a linear scheduler (0.9 decay ratio) is added. Test results are reported with the best validation scores.

[5]https://github.com/huggingface/transformers

Model: RoBERTa-base	**In-Domain & In-Target** (UB.)			**Single Corpus Training**		
Test Set	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
SemEval-16 Task-6 A	0.6849	0.6755	0.7169	0.6849	0.6755	0.7169
SemEval-16 Task-6 B	-	-	-	0.4134	0.5132	0.4389
P-Stance	0.6344	0.6435	0.6288	0.3454	0.4840	0.3980
VAST	0.7375	0.7499	0.7373	0.4079	0.4215	0.4140
Tweet-COVID	0.7474	0.7534	0.7483	0.3579	0.4334	0.4032
Model: RoBERTa-base	**Only Enriched Train Set**			**Adding Enriched Train Set**		
Test Set	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
SemEval-16 Task-6 A	0.6862	0.6774	0.7095	0.7047	0.6912	0.7355
SemEval-16 Task-6 B	0.6439	0.6493	0.6409	0.6885	0.6994	0.7010
P-Stance	0.4782	0.5175	0.4872	0.5003	0.5152	0.5007
VAST	0.6278	0.6488	0.6426	0.6346	0.6783	0.6462
Tweet-COVID	0.5202	0.5624	0.5349	0.5599	0.5821	0.5752

Table 6: Results of the 3-class stance classification on multiple corpora. Macro-averaged F1, Precision, and Recall scores are reported. *UB.* denotes the upper bound result from in-domain and in-target training on each corpus. Results of 2-class macro-averaged scores are shown in Appendix Table 10. Some examples of model prediction are shown in Appendix Table 12.

As previous work (Mohammad et al., 2016; Allaway and Mckeown, 2020), we adopt the macro-averaged F1, Precision, and Recall scores as evaluation metrics. The macro-averaged scheme weighs each of the classes equally and is not influenced by the imbalanced sample number of each class.

5.5 Experimental Results

Since we train the model on a single corpus (2.9k samples), testing it on multiple out-of-domain data and various unseen targets poses a challenging task. As shown in Table 6, compared with in-domain and in-target training on each corpus (which serves as the upper bound for external testing), scores of single-corpus training become much lower, and F1, precision, and recall are all affected significantly. This indicates that the original data only enable the model to achieve reasonable results on in-domain samples and existing targets. In contrast, training with the strategically-enriched annotation (1,500 paired samples) improves the performance substantially and consistently: on the four external test sets, the *RoBERTa-base* model has achieved at least 38% relative improvement of 3-class labeling. This demonstrates that the model learns more general and domain-invariant features which are useful across different stance detection corpora. Moreover, merging the original data and the enriched set brings further improvement, where at least 45% relative improvement of 3-class labeling is observed.

For extensive comparison with previous work (Mohammad et al., 2016; Li and Caragea, 2021), aside from the 3-class calculation, we report 2-class macro-averaged scores (i.e., *Favor*, *Against*), where the *None* label is used during training, but discarded in evaluation. As shown in Table 10, training with enriched data also provides a significant improvement (at least 48% relative gain), and state-of-the-art cross-target results.

5.6 Analysis on Target Dependency

Previous work found that strong baselines often solely rely on the input text but neglect the specified targets (Ghosh et al., 2019; Kaushal et al., 2021), resulting in poor performance on diverse and unseen targets. Since our model (trained with enriched set) is expected to show better target dependency, we envision that on label-balanced test sets, the distributions of predictions with or without specified targets shall be pretty distinct.

Here we conduct a quantitative analysis based on KL divergence as in Equation 2. Given $Q(x)$ is the prediction solely on the input text, and $P(x)$ is conditioned on the specified target, we calculate their KL divergence on the whole test set $\mathcal{X}$ to measure their similarity.

$$\mathrm{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log(\frac{P(x)}{Q(x)}) \quad (2)$$

As shown in Figure 3, compared with training on the original set, adding the enriched data results in larger KL divergence values. This empirically shows that model's prediction depends more on the specified targets than the base model.

Test Set	Single Corpus Training			Adding Enriched Train Set		
	ATAE	PoE	BERTweet	ATAE	PoE	BERTweet
SemEval16 Task-6 A	0.5604	0.6095	0.6833	0.5676	0.6507	0.7110
SemEval16 Task-6 B	0.2744	0.4853	0.5488	0.3297	0.6623	0.6533
P-Stance	0.3263	0.3720	0.4255	0.3401	0.4939	0.4763
VAST	0.3227	0.3656	0.3830	0.3703	0.6220	0.5757
Tweet-COVID	0.3146	0.3807	0.4707	0.4459	0.5048	0.5658

Table 7: Various model performance of the 3-class stance classification on multiple corpora. Macro-averaged F1 scores are reported. Results of the 2-class macro-averaged scores are shown in Appendix Table 11.

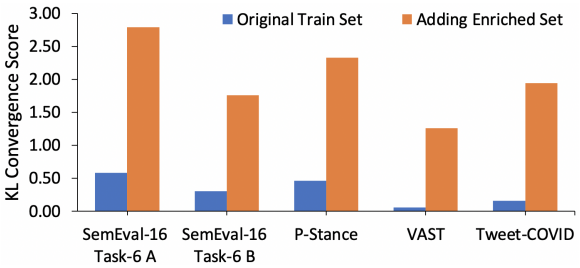

Figure 3: KL divergence comparison of predictions with targets and without targets.

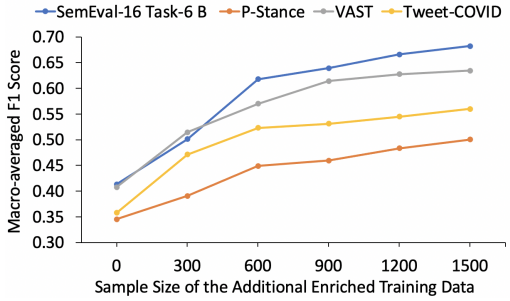

Figure 4: Results on different enriched annotation sizes. Y axis denotes the macro-averaged F1 Score calculated on 3-class prediction, and X axis denotes the additional sample size.

5.7 Analysis on Enriched Annotation Size

We then take an assessment on different sizes of the enriched annotation. As shown in Figure 4, compared with training on the original sample set, the performance on out-of-domain and cross-target evaluation can be boosted with 600 enriched samples. We speculate that by leveraging the prior knowledge of pre-trained language backbones, models can learn the general features effectively and efficiently from the enriched data. This demonstrates one advantage of following a linguistic framework for strategic annotation, where models can obtain substantial gain upon limited annotation cost.

5.8 Effectiveness across Model Architectures

We further conduct experiments with other strong baselines in different model architectures and designs: (1) **ATAE** (Wang et al., 2016): an LSTM-based model that extracts target-specific features via an attention mechanism. (2) **BERTweet** (Nguyen et al., 2020): a BERT-based language backbone that is specially pre-trained on tweet data. (3) **Product-of-Expert** (PoE) (Clark et al., 2019): a de-biasing method that reweights the learned probabilities upon a bias-only model. We train and evaluate these models following the same experimental setting described in Section 5, and their full implementation details are shown in Appendix Table 8. As shown in Table 7 and Table 11, the results in most aspects are improved substantially after adding the enriched data, which shows that the strategic augmentation is effective on various architectures. In addition, BERT-based models (e.g., *BERTweet*, *PoE*) show a larger performance gain than the *ATAE*, as they leverage prior language knowledge from pre-training. The comparable result of *RoBERTa-base*, *BERTweet*, and *PoE* shows various backbones can learn domain-invariant features from the enriched data, demonstrating the effectiveness of our adversarial multi-target annotation strategy (Section 4).

6 Related Work

Linguistic Studies of Stancetaking Stance analysis plays an essential role in measuring public opinion on mass media, conversations, and online networking platforms, particularly related to social, religious, and political issues (Agha, 2003; Kiesling, 2022). There are many qualitative studies conducted on various resources such as news interviews (Haddington et al., 2004), Twitter and Facebook posts (Simaki et al., 2018), narrative articles (Bohmann and Ahlers, 2022), and online forums (Kiesling et al., 2018). Recent works also perform in-depth content analyses on social media images to understand how politicians use images to express ideological rhetoric (Xi et al., 2020). To conduct analyses of stancetaking in a well-formulated man-

ner, many linguistic studies explore the explanation for the semantics, pragmatics, syntactic distribution of lexical items, discourse markers, and syntactic construction across languages (Biber and Finegan, 1988; Haddington et al., 2004; Lempert, 2008; Jaffe et al., 2009; Du Bois and Kärkkäinen, 2012). From the perspective of social pragmatics, stancetaking on social networks is regarded as a dynamic and dialogic activity where participants are actively engaged in virtual interactions (Du Bois, 2007). Tweets are not viewed as stagnant posts, instead, they become back-and-forth interactions enacted by retweets and comments (Chiluwa and Ifukor, 2015; Evans, 2016).

Computational Stance Detection Computational stance detection is commonly formulated as a target-specified classification problem. Datasets for stance detection are usually collected from online networking platforms where large-scale user-generated content is available (Mohammad et al., 2016; Li et al., 2021; Allaway and Mckeown, 2020; Glandt et al., 2021). Support vector machines (SVM) with manually-engineered features served as the earlier strong baseline (Mohammad et al., 2016). Then various deep learning techniques such as recurrent neural networks (RNNs) (Zarrella and Marsh, 2016), convolutional neural networks (CNNs) (Vijayaraghavan et al., 2016), and attention mechanism (Augenstein et al., 2016; Du et al., 2017; Zhou et al., 2017; Liu and Na, 2018) are applied for better feature extraction. Recently, the Transformer networks (Vaswani et al., 2017), especially language backbones (Devlin et al., 2019; Liu et al., 2019), boosted the performance on stance detection benchmarks (Ghosh et al., 2019; Li and Caragea, 2021). However, it has been observed that current strong baselines relied heavily on superficial features in existing data, and showed poor performance on unseen-target and out-of-domain evaluation (Kaushal et al., 2021), and recent work proposed de-biasing methods (Clark et al., 2019; Karimi Mahabadi et al., 2020) and introduced multi-task learning (Yuan et al., 2022).

7 Conclusions

In this work, we revisited the challenges of computational stance detection from a linguistic perspective. We expanded a generic linguistic framework the "stance triangle", with the relationship and labels of explicit and implicit objects, and characterized various fashions of how humans express their stances. We then followed the framework to strategically enrich the annotation of one benchmarked single-domain corpus. Experimental results showed that the enriched data significantly improve the performance on out-of-domain and cross-target evaluation, and guiding computational stance detection with expanded stance triangle framework can encourage models to learn more general and domain-invariant features. Moreover, our framework paves the way for future research such as assessing the explainability of data-driven models.

Limitations

All samples used in this work are in English, thus to apply the model to other languages, it will require training data on the specified language or using multilingual language backbones. Moreover, we are aware that it remains an open problem to mitigate biases in human stancetaking. Of course, current models and laboratory experiments are always limited in this or similar ways. We do not foresee any unethical uses of our proposed methods or their underlying tools, but hope that it will contribute to reducing incorrect system outputs.

Ethics and Impact Statement

We acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. All data used in this work are collected from existing published NLP studies. Following previous work, the annotated corpora are only for academic research purposes and should not be used outside of academic research contexts. Our proposed framework and methodology in general do not create a direct societal consequence and are intended to be used to prevent data-driven models from over-fitting domain-dependent and potentially-biased features.

Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R), A*STAR, Singapore, and DSO National Laboratories, Singapore. We thank Yizheng Huang and Stella Yin for preliminary data-related discussions, and Siti Umairah Md Salleh, Siti Maryam Binte Ahmad Subaidi, Nabilah Binte Md Johan, and Jia Yi Chan for linguistic resource construction. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Asif Agha. 2003. The social life of cultural value. *Language & communication*, 23(3-4):231–273.

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.

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.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse processes*, 11(1):1–34.

Axel Bohmann and Wiebke Ahlers. 2022. Stance in narration: Finding structure in complex sociolinguistic variation. *Journal of Sociolinguistics*, 26(1):65–83.

Innocent Chiluwa and Presley Ifukor. 2015. 'war against our children': Stance and evaluation in# bringbackourgirls campaign discourse on twitter and facebook. *Discourse & Society*, 26(3):267–296.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China. Association for Computational Linguistics.

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,*, pages 4171–4186.

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

John W Du Bois. 2007. The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, 164(3):139–182.

John W Du Bois and Elise Kärkkäinen. 2012. Taking a stance on emotion: Affect, sequence, and intersubjectivity in dialogic interaction. *Text & Talk*, 32(4):433–451.

Ash Evans. 2016. Stance and identity in twitter hashtags. *Language@ internet*, 13(1).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: a comparative study. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 75–87. Springer.

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.

Pentti Haddington et al. 2004. Stance taking in news interviews. *SKY Journal of Linguistics*, 17:101–142.

Alexandra Jaffe et al. 2009. *Stance: sociolinguistic perspectives*. Oup Usa.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. twt–wt: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889.

Scott F Kiesling. 2022. Stance and stancetaking. *Annual Review of Linguistics*, 8:409–426.

Scott F. Kiesling, Umashanthi Pavalanathan, Jim Fitzpatrick, Xiaochuang Han, and Jacob Eisenstein. 2018. Interactional stancetaking in online forums. *Computational Linguistics*, 44(4):683–718.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

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

Michael Lempert. 2008. The poetics of stance: Textmetricality, epistemicity, interaction. *Language in Society*, 37(4):569–592.

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.

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.

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*.

Zhengyuan Liu and Jin-Cheon Na. 2018. Aspect-based sentiment analysis of nuclear energy tweets with attentive deep neural network. In *ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings 20*, pages 99–111. Springer.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Vasiliki Simaki, Panagiotis Simakis, Carita Paradis, and Andreas Kerren. 2018. Detection of stance-related characteristics in social media text. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 1–7.

John S Uebersax. 1982. A generalized kappa coefficient. *Educational and Psychological Measurement*, 42(1):181–183.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

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. *arXiv preprint arXiv:1606.05694*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the political ideology of legislators from social media images. In *Proceedings of the international aaai conference on web and social media*, volume 14, pages 726–737.

Jianhua Yuan, Yanyan Zhao, Yanyue Lu, and Bing Qin. 2022. Ssr: Utilizing simplified stance reasoning process for robust stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6846–6858.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784*.

Yiwei Zhou, Alexandra I Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*, pages 18–32. Springer.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

Environment Details	
GPU Model	Single Tesla A100 with 40 GB memory; CUDA version 10.1.
Library Version	Pytorch==1.8.1; Transformers==4.8.2.
Computational Cost	Average 1.5 hours training time for one round. Average 3 rounds for each reported result (calculating the mean of the result scores).
Stance Classification	**Experimental Configuration**
Corpus	The datasets we used for training and evaluation are from published works (Mohammad et al., 2016; Allaway and Mckeown, 2020; Li et al., 2021; Glandt et al., 2021) with the Creative Commons Attribution 4.0 International license.
Pre-Processing	All samples are in English, and only for research use. Upper-case, special tokens, and hashtags are retained.
RoBERTa-base	RoBERTa-base (Liu et al., 2019) Base Model: Transformer (12-layer, 768-hidden, 12-heads, 125M parameters). Learning Rate: 2e-5, AdamW Optimizer, Linear Scheduler: 0.9.
BERTweet	BERTweet-base (Nguyen et al., 2020) Base Model: Transformer (12-layer, 768-hidden, 12-heads, 130M parameters). Learning Rate: 2e-5, AdamW Optimizer, Linear Scheduler: 0.9.
ATAE	ATAE-LSTM (Wang et al., 2016) Base Model: Bi-LSTM (2-layer Bi-directional LSTM, hidden dimension is 300, linear calculation of target-level attention, 15M parameters). Learning Rate: 3e-4, Adam Optimizer, Word Embedding: GloVe-840B.
Product-of-Expert (PoE)	Product-of-Expert (Clark et al., 2019) Base Model: RoBERTa (12-layer, 768-hidden, 12-heads, 125M parameters). The bias-only model is trained on the tweet text without specified targets. Learning Rate: 2e-5, AdamW Optimizer, Linear Scheduler: 0.9.

Table 8: Details of the experimental environment and the hyper-parameter setting.

Annotation Enrichment	
Original Dataset	The dataset we used for annotation enrichment is from published work (Mohammad et al., 2016) with the Creative Commons Attribution 4.0 International license.
Pre-Processing	All samples are in English, and only for research use. Upper-case, special tokens, and hashtags are retained. The original sample size is 2,914. We filter out the samples where the specified target is explicitly mentioned in the text. Next, to obtain explicit objects in an extractive manner, we apply an off-the-shelf constituency parsing model, and collect all noun phrases in the constituency tree.
Annotator Information	Four linguistic experts who are employed as full-time staff for natural language processing research participate in the task. Their major language is English. The gender distribution covers female and male.
Annotation Instruction	Each sample is a piece of text and a specified target, which forms one row in an excel file. Participants are asked to annotate the stance of the text author toward the specified target. The stance label is 3-class: *Favor*, *Against*, and *None*. We introduce the stance detection task, and show some examples to all participants as preparation. In addition, there are two automatically-generated attributes: explicit mention and label alignment, which can be calculated after the manual stance labeling.
Data Statistics	The enriched set contains 1,500 sample pairs, where each sample has annotation upon two different targets, and the adversarial pair size is 1.1k.
Data Availability	Upon acceptance, following the previous published work (Mohammad et al., 2016), the data can be accessed with the Creative Commons Attribution 4.0 International license, and only for research use.

Table 9: Details of the strategically-enriched data annotation.

Model: RoBERTa-base	**In-Domain & In-Target** (UB.)			**Single Corpus Training**		
Test Set	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
SemEval-16 Task-6 A	0.7023	0.7047	0.7318	0.7023	0.7047	0.7318
SemEval-16 Task-6 B	-	-	-	0.3143	0.5126	0.2814
P-Stance	0.7745	0.7538	0.7977	0.4436	0.6597	0.5118
VAST	0.6661	0.6637	0.6850	0.3905	0.4192	0.3906
Tweet-COVID	0.7244	0.7057	0.7500	0.2575	0.4377	0.1966
Model: RoBERTa-base	**Only Enriched Train Set**			**Adding Enriched Train Set**		
Test Set	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
SemEval-16 Task-6 A	0.7088	0.7029	0.7360	0.7264	0.7286	0.7468
SemEval-16 Task-6 B	0.6099	0.6337	0.5883	0.6463	0.7342	0.5956
P-Stance	0.6764	0.6704	0.7054	0.6844	0.6689	0.7027
VAST	0.5798	0.6437	0.5496	0.5826	0.7167	0.4908
Tweet-COVID	0.4184	0.5043	0.4147	0.4579	0.5595	0.4146

Table 10: Results of the 2-class stance classification on multiple corpora. Macro-averaged F1, Precision, and Recall scores are reported. *UB.* denotes the upper bound result from in-domain and in-target training on each corpus. Some examples of model prediction are shown in Appendix Table 12.

	Single Corpus Training			**Adding Enriched Train Set**		
Test Set	**ATAE**	**PoE**	**BERTweet**	**ATAE**	**PoE**	**BERTweet**
SemEval16 Task-6 A	0.6207	0.6531	0.7117	0.6266	0.6774	0.7347
SemEval16 Task-6 B	0.1743	0.4412	0.5208	0.2696	0.6051	0.6083
P-Stance	0.4129	0.5173	0.5742	0.4742	0.6831	0.6463
VAST	0.3134	0.4178	0.3604	0.3388	0.5532	0.5128
Tweet-COVID	0.2362	0.3124	0.4151	0.3784	0.3853	0.4951

Table 11: Various model performance of the 2-class stance classification on multiple corpora. Macro-averaged F1 scores are reported.

Multi-Target Prediction Examples		
Text:Considering the fact that Bush was a president of this country, I do not see it a joke that Trump is running ! #Election2016		
Given Target: "Bush"	**RoBERTa-base**: Against ✔	**Enhanced Model**: Against ✔
Given Target: "Donald Trump"	**RoBERTa-base**: Against ✗	**Enhanced Model**: Favor ✔
Text: If @SpeakerPelosi wants to keep her job, she will fix this election as we all know its rigged. First, Tom Perez must go... and Bernie is our president. How can a news station air results when the primary that didnt even happen yet?		
Given Target: "Pelosi"	**RoBERTa-base**: Against ✔	**Enhanced Model**: Against ✔
Given Target: "Bernie Sanders"	**RoBERTa-base**: Against ✗	**Enhanced Model**: Favor ✔
Text: Why not? This protects both the officer and the civilian and it keeps things transparent. Then it would not be simply a matter of opinion when things go awry. It will be on videotape. BUT how much will it cost to store all this data and for how long?		
Given Target: "bodycamera"	**RoBERTa-base**: Against ✗	**Enhanced Model**: Favor ✔
Given Target: "videotape"	**RoBERTa-base**: Against ✗	**Enhanced Model**: Favor ✔
Text: Why can't people take this virus seriously and wear a damn mask? I can't comprehend the childish behavior of some people who refuse to wear one. I just can't.		
Given Target: "wear face masks"	**RoBERTa-base**: None ✗	**Enhanced Model**: Favor ✔
Given Target: "refuse to wear one"	**RoBERTa-base**: None ✗	**Enhanced Model**: Against ✔

Table 12: Examples of the prediction with and without enriched data annotation. Correct and incorrect predictions are indicated with the ✔ and ✗ symbol, respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 9

☑ A3. Do the abstract and introduction summarize the paper's main claims?
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 5

☑ B1. Did you cite the creators of artifacts you used?
Section 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 5

☑ 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)?
Section 5, Section 9

☑ 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?
Section 5, Section 9

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5

☑ 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 5

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 5

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 5

☑ 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 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4, Table 9

☑ 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 4, Table 9

☑ 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)?
Section 4, Table 9

☐ 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?
Not applicable. No data collection. Use published open dataset.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No data collection. Use published open dataset.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 4, Table 9