P-Stance: A Large Dataset for Stance Detection in Political Domain

Yingjie Li[♣] **Tiberiu Sosea**[♣] **Aditya Sawant**[♣] **Ajith Jayaraman Nair**[♣]
Diana Inkpen[♢] **Cornelia Caragea**[♣]

[♣]Computer Science, University of Illinois at Chicago
[♢]Electrical Engineering and Computer Science, University of Ottawa
[♣]{yli300,tsosea2,asawan8,anair34,cornelia}@uic.edu
[♢]{diana.inkpen}@uottawa.ca

Abstract

Stance detection determines whether the author of a text is in favor of, against or neutral to a specific target and provides valuable insights into important events such as presidential election. However, progress on stance detection has been hampered by the absence of large annotated datasets. In this paper, we present P-STANCE, a large stance detection dataset in the political domain, which contains 21,574 labeled tweets. We provide a detailed description of the newly created dataset and develop deep learning models on it. Our best model achieves a macro-average F1-score of 80.53%, which we improve further by using semi-supervised learning. Moreover, our P-STANCE dataset can facilitate research in the fields of cross-domain stance detection such as cross-target stance detection where a classifier is adapted from a different but related target. We publicly release our dataset and code.[1]

1 Introduction

Nowadays, people often express their stances toward specific targets (e.g., political events or figures, religion, or abortion) on social media. These opinions can provide valuable insights into important events, e.g., presidential election. The goal of the stance detection task is to determine whether the author of a piece of text is in favor of, against, or neutral toward a specific target (Mohammad et al., 2016b; Küçük and Can, 2020; ALDayel and Magdy, 2021). Twitter as a social platform has produced a large quantity of user-generated content, which has become a rich source for mining useful information about various topics such as presidential election. Political figures, who usually receive considerable attention and involve themselves in a large number of political events, are great targets to study stance detection. Therefore, detecting the stance expressed toward political figures on Twitter has drawn a lot of attention in the NLP community (Mohammad et al., 2016a; Sobhani et al., 2017; Darwish et al., 2017).

Even though stance detection has received a lot of attention, the annotated data are usually limited, which poses strong challenges to supervised models. Moreover, a limitation of existing datasets is that explicit mentions of targets and surface-level lexical cues that may expose the stance can be widely observed in the data (Mohammad et al., 2016a; Sobhani et al., 2017; Swami et al., 2018; Darwish et al., 2018; Conforti et al., 2020b; Lai et al., 2020), which means a model can detect the stance without extracting effective representations for the meanings of sentences (i.e., their lexical and compositional semantics). Another limitation of existing datasets, especially the datasets built on social media, is that the average length of tweets is short, which indicates that the data in these previous datasets are less informative and thus the stance can be detected more easily.

In an effort to minimize these drawbacks, we present P-STANCE, a dataset for stance detection whose primary goal is to bridge these gaps by making it possible to run large-scale evaluations that require a deeper semantic understanding. This large annotated dataset is composed of 21,574 English tweets in the political domain and each tweet is annotated with a stance toward one of three different targets: "Donald Trump," "Joe Biden," and "Bernie Sanders." Examples from our dataset and their stance labels are shown in Table 1.

The main motivation of building this dataset is to provide a new benchmark for in-target stance detection where a classifier is trained and validated on the same target. However, we show additional interest in constructing a large corpus to facilitate research on cross-target stance detection where a classifier is adapted from different but related target.

[1] https://github.com/chuchun8/PStance

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2355–2365

Target	Tweet	Stance
Donald Trump	I agree, but not convinced Barr has all the evidence for his opinion. Although, I have zero worries about POTUS being re-elected, if the evidence is compelling enough on top of the tyrannical Covid lockdowns, Im hopeful more people will wake up. #GiantRedPill #Trump	Favor
Donald Trump	Take my kids, for example. At least, I'm TOLD they're my kids. No proof. Don Jr, Ivanka and Eric were all born to an immigrant woman who WASN'T a US Citizen when they were born. They shouldn't have US Citizenship. DEPORT THEM ALL! #Trump	Against
Bernie Sanders	Air borne illnesses will only become more common with climate change. We need to immediately address this and fight for Medicare for All or this could be the new normal. #BernieSanders	Favor
Bernie Sanders	A meat tax? Paying off all medical bills? I think #bernie has truly gone off the deep end of the pander cliff. None of these socialists insane, pie-in-the-sky policies would EVER work, or even come in to fruition yet people continue to fall for it. Unbelievable. #foxandfriends	Against
Joe Biden	Robyn Seniors, National HBCU Students for Biden Co-Chair and a @FAMU_1887 student, says that she's thankful that a "woman will be Vice President in a Biden administration."	Favor
Joe Biden	The Ukrainians are smarter than our own democratic party! Shoot, my Dogs are smarter than our own democratic party!! #ImpeachmentHoax #NoQuidProQuo #Biden	Against

Table 1: Examples from our P-STANCE dataset.

More interestingly, P-Stance enables a new task in stance detection, which is cross-topic stance detection where a classifier is adapted from the same target but with different topics in the past. These tasks, which use labeled training data of a source target and aim to train a model that generalizes well to a destination target with a shifted distribution, hold great practical value.

Our contributions include the following: 1) We present P-STANCE, a large dataset for stance detection composed of 21,574 tweets sampled from over 2.8 million tweets collected from Twitter. P-STANCE is more than three times larger than the previous benchmark (Mohammad et al., 2016a) and brings additional challenges such as linguistic complexities. We provide a detailed description and a comprehensive analysis of this dataset; 2) We conduct experiments on the proposed P-STANCE dataset and establish a strong baseline based on BERTweet (Nguyen et al., 2020). BERTweet achieves a macro-average F1-score of 80.53%, which we improve further by using semi-supervised learning; 3) The union of P-STANCE and previous benchmark datasets provides more opportunities for studying other stance detection tasks, e.g., cross-target stance detection and cross-topic stance detection.

2 Related Work

The most common stance detection task on social media is target-specific stance detection (ALDayel and Magdy, 2021) which aims to identify the stance toward a set of figures or topics (Hasan and Ng, 2014; Mohammad et al., 2016a; Xu et al., 2016; Taulé et al., 2017; Swami et al., 2018; Zotova et al., 2020; Conforti et al., 2020b; Lai et al., 2020; Vamvas and Sennrich, 2020; Conforti et al., 2020a). Besides target-specific stance detection, multi-target stance detection (Sobhani et al., 2017; Darwish et al., 2017; Li and Caragea, 2021a), and claim-based stance detection (Qazvinian et al., 2011; Derczynski et al., 2015; Ferreira and Vlachos, 2016; Bar-Haim et al., 2017; Rao and Pomerleau, 2017; Derczynski et al., 2017; Gorrell et al., 2019) are other popular trends of stance detection. Multi-target stance detection aims to jointly identify the stance toward two or more targets in the same text. Unlike the target-specific stance detection and multi-target stance detection where the target is usually a prominent figure or topic, in claim-based stance detection the target is a claim, which could be an article headline or a rumor's post.

Interestingly, despite substantial progress on stance detection, large-scale annotated datasets are limited. We compare our P-STANCE dataset with some existing stance detection datasets in Table 2. We can observe that the sizes of existing stance detection datasets are smaller than ours except for the WT-WT dataset (Conforti et al., 2020b) in the financial domain. However, the average tweet length of WT-WT is much shorter when compared with our P-STANCE. Moreover, more explicit mentions of targets and lexical cues of stance appear in the sentences of WT-WT dataset. In our work, we focus on the political domain and our P-STANCE, which contains much longer sentences and less surface-level lexical cues, can serve as a new challenging benchmark for stance detection tasks.

Authors	Target(s)	Source	Type	Size
Mohammad et al. (2016a)	Atheism, Climate change is a real concern, Feminist movement, Hillary Clinton, Legalization of abortion, Donald Trump	Twitter	Target-specific	4,870
Ferreira and Vlachos (2016)	Various claims	News articles	Claim-based	2,595
Sobhani et al. (2017)	Trump-Clinton, Trump-Cruz, Clinton-Sanders	Twitter	Multi-target	4,455
Derczynski et al. (2017)	Various claims	Twitter	Claim-based	5,568
Swami et al. (2018)	Demonetisation in India in 2016	Twitter	Target-specific	3,545
Gorrell et al. (2019)	Various claims	Twitter, Reddit	Claim-based	8,574
Conforti et al. (2020b)	Merger of companies: Cigna-Express Scripts, Aetna-Humana, CVS-Aetna, Anthem-Cigna, Disney-Fox	Twitter	Target-specific	51,284
Conforti et al. (2020a)	Merger of companies: Cigna-Express Scripts, Aetna-Humana, CVS-Aetna, Anthem-Cigna	News articles	Target-specific	3,291
P-STANCE	Donald Trump, Joe Biden, Bernie Sanders	Twitter	Target-specific	21,574

Table 2: Comparison of English stance detection datasets.

Different from classifying the stance detection tasks by target type (i.e., one specific target, multiple targets, or a claim), we can also categorize the stance detection as in-target and cross-target stance detection by the training setting. Most previous works focused on the in-target stance detection where a classifier is trained and validated on the same target (Mohammad et al., 2016b; Zarrella and Marsh, 2016; Wei et al., 2016; Vijayaraghavan et al., 2016; Du et al., 2017; Sun et al., 2018; Wei et al., 2018; Li and Caragea, 2019, 2021b). However, sufficient annotated data are usually hard to obtain and conventional models on stance detection perform poorly on generalizing to the data of new targets, which motivates the studies of cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Wei and Mao, 2019; Zhang et al., 2020). Most previous studies evaluated the cross-target models on the SemEval-2016 dataset (Mohammad et al., 2016a), which is a small dataset and thus may make the conclusions less convincing.

In this paper, we show that our P-STANCE dataset can be also used to evaluate the model performance of cross-target stance detection and provides opportunities for exploring more cross-target tasks by interacting with previous SemEval-2016 (Mohammad et al., 2016a) and Multi-Target stance datasets (Sobhani et al., 2017). In addition, P-STANCE enables the exploration of large-scale deep learning models including pre-trained language models, e.g., BERT (Devlin et al., 2019) and BERTweet (Nguyen et al., 2020). We fine-tune the BERT and BERTweet models on our dataset and compare them with other strong baselines.

3 Building the Dataset

In this section, we detail the creation and the particularities of P-STANCE, our large political stance detection dataset composed of 21,574 tweets collected during the 2020 U.S. presidential election.

3.1 Data Collection

We collected tweets using the Twitter streaming API. Similar to prior works (Mohammad et al., 2016a; Sobhani et al., 2017) that target presidential candidates, we focus our attention on three political figures[2] in the presidential race of 2020: "Donald Trump," "Joe Biden," and "Bernie Sanders." We used a set of query hashtags as seeds to collect target-related tweets, which can be categorized as *favor hashtags*, *against hashtags* and *neutral hashtags* (Mohammad et al., 2016a). We show examples of these query hashtags in Table 3. In total, we gathered around 2.8 million tweets for all three targets combined.

3.2 Preprocessing

To ensure the quality of this dataset, we performed several preprocessing steps: **1)** We removed tweets with less than 10, or more than 128 words. According to our observations, tweets with less than 10 words are either too easy for detecting the stance or too noisy, and tweets with more than 128 words usually contain duplicate expressions. **2)** We removed duplicates and retweets. Twitter data are noisy not only due to the creative spellings, slang

[2]We also tried to collect tweets about the woman politician Kamala Harris. However, we were unable to collect enough data about Harris. We will look into this in our future work.

Target	Favor Hashtag	Against Hashtag	Neutral Hashtag
Trump	#Trump2020LandSlide	#TrumpCrimeFamily	#DonaldTrump #Republican
Biden	#BidenForPresident	#SleepyJoe	#JoeBiden #Democrats
Sanders	#BernieWon	#NeverBernie	#BernieSanders #Sanders

Table 3: Examples of query hashtags.

	Trump	Biden	Sanders
#Raw collection	1,730K	429K	654K
#After preprocessing	1,221K	300K	465K

Table 4: Number of unlabeled tweets before and after preprocessing.

Setup	Trump	Biden	Sanders	Average
3-class	0.62	0.60	0.59	0.60
2-class	0.86	0.81	0.76	0.81

Table 5: Krippendorff's alpha measure of annotator agreement in 3-class and 2-class scenarios.

and URLs, but also because of the duplicate tweets. Since these duplicate data reduce our ability to build reliable models, we need to clean the dataset by removing duplicates. **3)** We kept only the tweets in English because our goal in this work is to build an English stance detection dataset. We leave multilingual stance detection as future work. After data preprocessing, the size of our corpus reduces to around 2 million examples. In Table 4, we show the number of tweets before and after preprocessing for each political figure. We will provide this large-scale repository of tweets (which we call P-STANCE-EXT) alongside P-STANCE, in hope that it will spur further research in the field of semi-supervised learning for stance detection. Finally, we sampled 10,000 tweets for each political figure, obtaining 30,000 tweets for annotation in total.

3.3 Data Annotation

We gathered stance annotations of three targets through the Amazon Mechanical Turk (AMT) crowdsourcing platform. The AMT workers were asked to annotate each tweet with "Favor," "Against," "None," or "I don't know." To ensure the annotation quality, we employed strict requirements for the annotators: **1)** Many completed tasks (>500); **2)** To reside in the USA; **3)** A high acceptance rate (>95%). Moreover, we ran the annotation process in several batches of 1000 examples. In each batch, we include 100 internally annotated examples to measure the quality of the annotators. If an annotator mislabels more than 25% of these examples, we discard the annotations of the worker completely, and relabel them. Interestingly, this process led to a considerable number of reannotations, amounting for more than 20% of the data. Each tweet was labeled by three random annotators, and disagreements in the labels were decided by the majority voting among the three annotators.

After obtaining the annotation results, we computed Krippendorff's alpha (Krippendorff, 2011) as the measure of inter-annotator agreement, as shown in Table 5. Tweets that were annotated with label "I don't know" after the majority voting were removed from the dataset. We observed that annotators had difficulties in reaching an agreement on tweets with label "None" and the average of Krippendorff's alpha values increases from 0.60 to 0.81 when we consider two classes: "Favor" and "Against". Similar to prior work (Vamvas and Sennrich, 2020), we removed the label "None" from the dataset in our experiments.

3.4 Quality Assurance and Challenges

Stance-exposing hashtags that may expose the stance directly, e.g., #NeverBernie, can be observed in the data. A model can detect the stance from these hashtags without extracting effective representations for the meanings of sentences, which makes stance detection easier. To remove the stance-exposing hashtags and ensure the data quality, we performed the following steps after the data annotation: **1)** We manually built a hashtag lexicon that contains stance-exposing hashtags for each target. Then we removed all hashtags that are appended at the end of a sentence if they are in the hashtag lexicon. The reason of only removing the appended hashtags is that a hashtag may serve as a constituent of a sentence, so it would introduce more noise if we simply remove all stance-exposing hashtags. **2)** To address the stance-exposing hashtag that is a constituent of a sentence, we replaced stance-exposing hashtags that contain the target name with a neutral hashtag, e.g., #NeverBernie → #Bernie. These steps ensure

		Trump	Biden	Sanders
Train	Favor	2,937	2,552	2,858
	Against	3,425	3,254	2,198
Val	Favor	365	328	350
	Against	430	417	284
Test	Favor	361	337	343
	Against	435	408	292
Total		7,953	7,296	6,325

Table 6: Label distribution across different targets for P-STANCE.

the high quality of our P-STANCE dataset.

In addition, P-STANCE is a challenging dataset for the following reasons: **1)** Targets in P-STANCE are referred to in a more implicit way. Consider the second example in Table 1, the target name only appears at the end of the sentence and it is hard to correctly identify the stance without any knowledge about the political figures mentioned in the content and background immigration policy. Similarly, for the third example, it is difficult to correctly identify the stance if the classifier fails to connect the target with relevant events, i.e., climate change or medicare for all residents. **2)** The average length of tweets in previous datasets is short, and there are more explicit mentions of targets and rich sentiment and emotion words that can easily reveal the stance toward the target. The average tweet length is 17 in Mohammad et al. (2016a), 21 in Sobhani et al. (2017) and 16 in Conforti et al. (2020b). However, our P-STANCE has a much longer average length of 30 and more implicit mentions of targets and context words, which indicates that our dataset is more difficult. In addition, P-STANCE covers more target-relevant events. These characteristics contribute to making P-STANCE a challenging dataset for stance detection.

3.5 Dataset Distribution

The final dataset contains 7,953 annotated tweets for "Donald Trump", 7,296 for "Joe Biden" and 6,325 for "Bernie Sanders", respectively. The label distribution of each target is shown in Table 6. Each tweet is annotated with a stance label "Favor" or "Against". We created the training, validation and testing sets following an 80/10/10 split. We note that P-STANCE is more than 3 times larger than the previous benchmark (Mohammad et al., 2016a).

4 Experimental Settings

In this section, we first introduce two benchmark datasets of stance detection in §4.1. The union of these datasets and our P-STANCE dataset provides opportunities for studying the cross-target stance detection (§5.2) and cross-topic stance detection (§5.3). Then we discuss the evaluation metrics in §4.2 and introduce the baseline methods in §4.3.

4.1 Existing Benchmark Datasets

SemEval-2016 (Mohammad et al., 2016a) and Multi-Target stance datasets (Sobhani et al., 2017) are two benchmark datasets in which political figures are chosen as the targets. SemEval-2016 contains six targets: "Atheism," "Climate Change is a Real Concern," "Feminist Movement," "Hillary Clinton," "Legalization of Abortion," and "Donald Trump." The dataset is annotated for detecting the stance toward a given target. The data distribution of SemEval-2016 is shown in Table 7.

Multi-Target stance dataset contains three sets of tweets corresponding to three target pairs: "Donald Trump and Hillary Clinton," "Donald Trump and Ted Cruz," "Hillary Clinton and Bernie Sanders" for 2016 U.S. presidential election. The task aims at detecting the stances toward two targets for each data. The data distribution of Multi-Target stance dataset is shown in Table 8. In the next section, we show how to perform various stance detection tasks with the union of these datasets and our P-STANCE dataset.

4.2 Evaluation Metrics

Similar to Mohammad et al. (2017) and Sobhani et al. (2017), F_{avg} and macro-average of F1-score (F_{macro}) are adopted to evaluate the performance of our baseline models. First, the F1-score of label "Favor" and "Against" is calculated as follows:

$$F_{favor} = \frac{2P_{favor}R_{favor}}{P_{favor} + R_{favor}} \quad (1)$$

$$F_{against} = \frac{2P_{against}R_{against}}{P_{against} + R_{against}} \quad (2)$$

where P and R are precision and recall, respectively. After that, the F_{avg} is calculated as:

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \quad (3)$$

We compute the F_{avg} for each target. F_{macro} is calculated by averaging the F_{avg} across all targets.

4.3 Baseline Methods

We run experiments with the following baselines.

Target	#Train	%Favor	%Against	%None	#Test	%Favor	%Against	%None
Atheism	513	17.93	59.26	22.81	220	14.54	72.73	12.73
Climate	395	53.67	3.80	42.53	169	72.78	6.51	20.71
Feminism	664	31.63	49.40	18.97	285	20.35	64.21	15.44
Hillary	689	17.13	57.04	25.83	295	15.25	58.31	26.44
Abortion	653	18.53	54.36	27.11	280	16.43	67.50	16.07
Trump	0	-	-	-	707	20.93	42.29	36.78

Table 7: Data distribution of SemEval-2016 dataset.

Target Pair	Total	Train	Dev	Test
Trump-Clinton	1,722	1,240	177	355
Trump-Cruz	1,317	922	132	263
Clinton-Sanders	1,366	957	137	272
Total	4,455	3,119	446	890

Table 8: Data distribution of Multi-Target dataset.

BiLSTM (Schuster and Paliwal, 1997): A BiLSTM model that takes tweets as inputs without considering the target information.

CNN (Kim, 2014): Similar to BiLSTM, the vanilla CNN only takes tweets as inputs and does not consider the target information.

TAN (Du et al., 2017): TAN is an attention-based LSTM model that extracts target specific features.

BiCE (Augenstein et al., 2016): A BiLSTM that uses conditional encoding for stance detection. The target information is first encoded by a BiLSTM, whose hidden representations are then used to initialize another BiLSTM with tweets as inputs. BiCE is also a strong baseline for cross-target stance detection.

CrossNet (Xu et al., 2018): CrossNet is another model for cross-target stance detection. It encodes the target and the tweet by using the same approach with BiCE and add an aspect attention layer to signal the core part of a stance-bearing input. CrossNet improves BiCE in many cross-target settings.

GCAE (Xue and Li, 2018): A CNN model that utilizes a gating mechanism to block target-unrelated information. GCAE is a strong baseline for aspect-based sentiment analysis and we apply it to our stance detection task.

PGCNN (Huang and Carley, 2018): Similar to GCAE, PGCNN is based on gated convolutional networks and encodes target information by generating target-sensitive filters.

BERT (Devlin et al., 2019): A pre-trained language model that predicts the stance by appending a linear classification layer to the hidden representation of $[CLS]$ token. We fine-tune the BERT-base on the stance detection task.

BERTweet (Nguyen et al., 2020): BERTweet is another pre-trained language model following the training procedure of RoBERTa (Liu et al., 2019). Similar to BERT, we fine-tune the pre-trained BERTweet to predict the stance by appending a linear classification layer to the hidden representation of the $[CLS]$ token. The pre-trained BERTweet model is fine-tuned under the PyTorch framework. The maximum sequence length is set to 128 and the batch size is 32. We use AdamW optimizer (Loshchilov and Hutter, 2019) and the learning rate is 2e-5.

5 Results

In this section, we present the set of experiments performed on various stance detection tasks on our dataset and show the results obtained by using the aforementioned baselines. Each result is the average of seven runs with different initializations.

5.1 In-Target Stance Detection

In-target stance detection is a stance detection task where a classifier is trained and validated on the same target. Most previous works adopt an "Ad-hoc" training strategy by training one model for each target and evaluate it on the test set of that target (i.e., we train three different models if there are three targets in the dataset). However, the model is more likely to predict the stance by following specific patterns without fully considering the target information and overfit. Therefore, to better evaluate the performance of baselines, we propose a "Merged" training strategy by training and validating a model on all targets and testing it on separate targets to be compared with the "Ad-hoc" setting.

Experimental results of these two different settings are shown in Table 9. First, we can observe that BERTweet performs best in both settings and significantly outperforms the second best results,

Method	Trump	Biden	Sanders	F_{macro}	Drop
Ad-hoc					
BiLSTM	76.92	77.95	69.75	74.87	-
CNN	76.80	77.22	71.40	75.14	-
TAN	77.10	77.64	71.60	75.45	-
BiCE	77.15	77.69	71.24	75.36	-
PGCNN	76.87	76.60	72.13	75.20	-
GCAE	78.96	77.95	71.82	76.24	-
BERT	78.28	78.70	72.45	76.48	-
BERTweet	**82.48**†	**81.02**†	**78.09**†	**80.53**	-
Merged					
BiLSTM	77.18	75.47	67.43	73.36	1.51
CNN	74.79	74.11	66.68	71.86	3.28
TAN	78.30	75.26	70.67	74.74	0.71
BiCE	77.67	75.69	69.37	74.24	1.12
PGCNN	77.36	74.96	70.29	74.20	1.00
GCAE	79.00	76.32	69.93	75.08	1.16
BERT	79.19	76.02	73.59	76.27	0.21
BERTweet	**83.81**†	**79.08**†	**77.75**†	**80.21**	0.32

Table 9: Comparison of different models on the P-STANCE dataset (%). †: BERTweet model improves the best baseline at $p < 0.05$ with paired t-test. F_{macro} is the average of all target pairs. "Drop" means performance decline between two training strategies for the same model. Bold scores are best overall.

Target	BiCE	CrossNet	BERTweet
P-STANCE dataset			
DT → JB	55.83	56.67	**58.88**
DT → BS	51.78	50.08	**56.50**†
JB → DT	58.16	60.43	**63.64**†
JB → BS	60.24	60.81	**67.04**†
BS → DT	51.41	52.99	**58.75**†
BS → JB	57.68	62.57	**72.99**†
DT, JB → BS	52.26	56.26	**69.99**†
DT, BS → JB	53.73	55.57	**68.64**†
JB, BS → DT	53.91	56.44	**66.01**†
P-STANCE → previous datasets			
DT → HC	36.12	**40.56**	34.48
DT → TC	59.37	59.40	**63.89**†
DT → BS	47.73	48.93	**51.00**†
JB → DT	48.90	49.77	**56.00**†
JB → HC	56.77	55.54	**57.55**
JB → TC	53.47	55.77	**62.45**†
JB → BS	48.11	48.96	**51.48**†
BS → DT	47.93	46.10	**49.96**
BS → HC	49.97	50.49	**52.81**
BS → TC	54.37	52.98	**56.91**†

Table 10: Comparison of different models for cross-target stance detection (%). The first half reports the cross-target results on our proposed P-STANCE dataset. The second half reports the cross-target results that are trained on the P-STANCE dataset and tested on the previous datasets. †: BERTweet model improves the best baseline at $p < 0.05$ with paired t-test. Bold scores are best overall.

demonstrating the effectiveness of this model. Second, performance drops can be observed on all models in the "Merged" setting and models (BiLSTM and CNN) that do not consider target information suffer the most severe drops, which means our proposed training strategy can serve as a better evaluation method to test whether the model learns target-specific representations. Moreover, we can observe that both BERTweet and BERT perform well and have the minimum performance drops compared with the other baselines, which demonstrates that self-attention mechanism can better capture target-specific representations.

5.2 Cross-Target Stance Detection

Despite substantial progress on the stance detection, sufficient annotated data are usually hard to obtain and conventional models on stance detection perform poorly on generalizing to the data of new targets, which motivates the studies of cross-target stance detection. The model of cross-target stance detection is first trained and validated on a source target, and then tested on a destination target. In this subsection, we show that our P-STANCE dataset can be also used to evaluate the model performance of cross-target stance detection and provides opportunities for exploring more cross-target tasks by interacting with previous SemEval-2016 and Multi-Target stance datasets.

We use five targets for our experiments: "Donald Trump" (DT), "Joe Biden" (JB), "Bernie Sanders" (BS), "Hillary Clinton" (HC), and "Ted Cruz" (TC). Experimental results of cross-target stance detection are shown in Table 10. For the first half of Table 10, only targets of P-STANCE dataset are used to evaluate the model performance. However, for the second half, targets of SemEval-2016 and Multi-Target datasets also serve as destination targets, which makes it a more challenging task since the target-related topics in 2016 are quite different from the ones in 2020. More specifically, we train and validate the model on a source target of P-STANCE dataset and test it on the data of a destination target, which is a combination of train, validation, and test sets of previous datasets. Note that we merge the data from SemEval-2016 and Multi-Target datasets if these two datasets share the same target, e.g., Hillary Clinton.

For the cross-target tasks only on the P-STANCE dataset, first, we can observe from the Table 10 that BERTweet achieves the best performance on all target configurations, demonstrating its effectiveness. Moreover, BERTweet shows greater improvement over the best baseline when training on the data of two targets. The reason is that BERTweet learns

more universal representations by leveraging the data from two targets. Second, we see that CrossNet outperforms BiCE on almost all target configurations, which is consistent with the observations of previous studies (Xu et al., 2018; Zhang et al., 2020). Third, we find that models achieve better performance on JB → BS and BS → JB. One potential explanation is that targets "Joe Biden" and "Bernie Sanders" are from the same party and thus share more similar topics.

For the second half of Table 10, we observe a significant drop in performance on all models, which verifies that it is more challenging to transfer the knowledge to a destination target with more diverse topics in the past. BERTweet still achieves the best performance on almost all target configurations, making it a highly competitive model for cross-target stance detection task. Interestingly, we can observe that both BiCE, CrossNet, and BERTweet show better performance on target "Ted Cruz." A possible reason is that the data of "Ted Cruz" contain more universal expressions and topics.

5.3 Cross-Topic Stance Detection

Obtaining sufficient annotated data of specific target from most recent past is challenging. However, sometimes historical annotated data of the same target are available. Therefore, motivated by a desire to improve the models' generalization ability to transfer knowledge from historical data, we come up with a new stance detection task, named cross-topic stance detection. Specifically, in this task, the model of cross-topic stance detection is first trained on the data of a target (e.g., Donald Trump) in 2016, and then validated and tested on the data of the same target in 2020. Note that the annotated data of year 2016 are the same with the data used in §5.2. The results are shown in Table 11. Since target "Joe Biden" is absent from the previous stance detection datasets, we use targets "Donald Trump" and "Bernie Sanders" for evaluation. We can observe that BERTweet still performs best on this task and the overall model performance of cross-topic stance detection is better than that of cross-target stance detection due to the use of the same target in evaluation stage. Moreover, we see that models perform relatively poorly on target "Bernie Sanders". One possible explanation is that some topics, e.g. healthcare and climate change, appear rarely in previous datasets.

Target	BiCE	CrossNet	BERTweet
DT → DT	58.60	59.41	**73.58**†
BS → BS	59.04	57.66	**66.48**†

Table 11: Comparison of different models for cross-topic stance detection (%). †: BERTweet model improves the best baseline at $p < 0.05$ with paired t-test. Bold scores are best overall.

5.4 Semi-Supervised Stance Detection

During elections, there is a considerable amount of data generated by users expressing their opinions about candidates, out of which only a small amount can be annotated and used for supervised stance detection. We explore the potential of the abundant unlabeled tweets and show that we can leverage them to improve the performance of our models. To this end, we turn to semi-supervised learning, and leverage techniques such as Uncertainty-aware Self-Training (UST).

UST (Mukherjee and Awadallah, 2020) is a semi-supervised approach which uses the standard teacher-student self-training framework, but adds a few powerful changes. Concretely, UST designs different techniques which leverage the uncertainty of the teacher model to select the unlabeled set of examples in each self-training iteration. First, we train our teacher model on the labeled examples. Next, we compute uncertainty estimates of our teacher model on the set of unlabeled examples by performing a few forward passes with dropout enabled. Finally, we incorporate the uncertainty estimates into our framework as follows: **1)** We use these estimates to select the examples for which the teacher is most or least confident about. **2)** We incorporate the teacher confidence in the student loss by penalizing the student's misclassified examples in which the teacher has high confidence. We use the BERTweet model as teacher and student.

We perform various experiments to show the benefits of using a large amount of unlabeled data from P-STANCE-EXT alongside our UST model. We carry out three barely supervised experiments with various number of examples in the training set. Specifically, we experiment with 30, 50, and 100 training examples. Moreover, we also consider an experiment using the whole training set to investigate the effect of the unlabeled examples when all the training data are available. We run experiments with different training sets, and report the F1-scores obtained on the entire testing set.

We show the results of our semi-supervised ex-

Method	Trump	Biden	Sanders	F_{macro}
BERTweet	82.48	81.02	78.09	80.53
UST-30	61.02	64.34	60.45	61.94
UST-50	68.42	73.24	66.12	69.26
UST-100	74.45	79.46	71.67	75.19
UST-ALL	**85.50**†	**82.22**†	**79.55**†	**82.42**

Table 12: Semi-supervised learning results. †: UST-ALL improves the BERTweet at $p < 0.05$ with paired t-test.

periments in Table 12 and make the following observations. First, UST-ALL significantly outperforms the BERTweet model by 1.89% in a macro-average F1-score when using both the labeled and unlabeled data in a semi-supervised manner. Second, with only 100 examples (2% of the available training examples), UST-100 stays within 1.6% F1-score of our best model that leverages the entire training set of target "Joe Biden." The results indicate that the benefit of using semi-supervised approaches is two-fold. On one hand, it enables impressive performance in scarce label scenarios, while on the other hand, it still brings gains in scenarios where considerable amounts of labeled data are readily available.

6 Conclusion

In this paper, we introduced P-STANCE, an English stance detection dataset in the political domain, which is larger and more challenging compared with previous datasets for stance detection. Composed of 21,574 tweets that were collected during the 2020 USA election, P-STANCE can serve as a new benchmark for stance detection and enable future research in other stance detection tasks, e.g., cross-target stance detection and cross-topic stance detection. Experimental results show that the BERTweet model significantly outperforms other strong baselines not only on in-target stance detection, but also on cross-target and cross-topic stance detection. Moreover, the performance of BERTweet can be further improved by using semi-supervised learning. Future work includes constructing another large dataset for a more challenging task, i.e., multi-target stance detection, and studying the multilingual stance detection with the union of P-STANCE and other multilingual datasets.

Acknowledgments

We thank the National Science Foundation and Amazon Web Services for support from grants IIS-1912887 and IIS-1903963 which supported the research and the computation in this study. We also thank our reviewers for their insightful comments.

References

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

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.

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.

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.

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.

Kareem Darwish, Walid Magdy, Afshin Rahimi, Timothy Baldwin, and Norah Abokhodair. 2018. Predicting online Islamophobic behavior after #ParisAttacks. *The Journal of Web Science*, 4(3):34–52.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. Hillary: What went viral during the 2016 US presidential election. In *9th International Conference on Social Informatics (SocInfo 2017)*, pages 143–161.

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.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Tomas Pariente Lobo, Anna Kolliakou, Robert Stewart, Sara-Jayne Terp, Geraldine

Wong, Christian Burger, Arkaitz Zubiaga, Rob Procter, and Maria Liakata. 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.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3988–3994.

William Ferreira and Andreas Vlachos. 2016. Emergent: A novel dataset for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168.

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.

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.

Binxuan Huang and Kathleen Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1091–1096.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

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

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

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

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6298–6304.

Yingjie Li and Cornelia Caragea. 2021a. A multi-task learning framework for multi-target stance detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Findings, ACL 2021, Online Event, August 2-4, 2021*. Association for Computational Linguistics.

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

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*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *LREC*.

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.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 21199–21212.

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.

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.

Delip Rao and Dean Pomerleau. 2017. Fake News Challenge.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409.

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

Mariona Taulé, Maria Antònia Martí, Francisco M. Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the task on stance and gender detection in tweets on Catalan independence. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, pages 157–177.

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

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.

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.

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 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, 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.

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.

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

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523.

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.

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.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1368–1375.